CS685: Data Mining

Analysing The Agricultural Patterns in India

Group 20

Abhishek Bhatia	170022	avishek@iitk.ac.in
Harshit Kumar	170293	hakumar@iitk.ac.in
Gosai Akash	170278	akashg@iitk.ac.in
Tanmay Anand	170751	tanand@iitk.ac.in

Indian Institute of Technology Kanpur

Computer Science And Engineering

Contents

1 Abstract

Today, India stands second in the worldwide ranking of farming sector output. Agriculture industry contributes about 17% to the Indian GDP and provides employment to more than half of the Indian population. India is not only a very large scale consumer of crops but also a very significant exporter. Despite agriculture being done on such a large scale, the processes are highly inefficient and often lack scientific analysis. Times and again, certain complex decisions have to be taken not only by the farmers but also by some agribusinesses. In this project, we aim to analyze the factors and complexities involving those decisions by studying the data associated with various variables concerning agriculture. We tried to find correlation between various factors such as rainfall, soil type, soil ptH, temperature, literacy rates, cost of production, cultivation area, crop price etc. and their consequent effects on the agriculture and crop patterns.

2 Introduction

Let's say, if a farmer wants to start cultivation in his little land or an agribusiness house wants to set up a plant in some district, we wonder if there's a way in which they can figure out what would be the most beneficial crop for production in terms of yield or in terms of production cost in their area. As we know, there are a number of factors affecting the crop production. We tried to find a correlation between the geographical factors, economical factors, cultivation practices for different districts of India and see if we can help building a crop recommendation structure. We also tried to incorporate factors like the literacy rate of the districts and see if it has correlation with the crop production of the local area. For simplicity, we have divided our analysis into two major categories: Nationwide analysis and Geographical analysis. Along with that, we have implemented a predictive model that gives us the information about the kind of crops that should be cultivated in various regions depending on the geographical conditions as well as administrative factors such as crop price and literacy rate. Using a recommendation system, we gave recommendations on what would be the suitable crops to grow in each of the state.

3 Objectives

As already described above, we have categorized our analysis into various parts. The main objectives can be majorly summarized as:

1. **Geographical Analysis on Crop Production**: We know that demographic factors like soil type, rainfall, humidity, soil pH, nitrogen levels, etc. has direct effect on the crops production. Not only this, we took certain unusual factors such as literacy rates. We did a district-wise analysis of these factors on agriculture patterns. In this, we studied the impact of rainfall on cultivation area, impacts of soil type and soil pH on various crops, impact of literacy rates on crop production, etc.

2. **Nation-wide Analysis of Crops**: Here, we studied the year-wise changes in crop production throughout the India, the percentage shares of various crops in India and tried to figure out the reasons for their changing trends, if any. For this, we have tried analysing the year-wise changes in crop price(averaged over districts), the crops production throughout the country, revenues of the crops, production and cultivation costs, earnings per area, etc.

3. **Predictive Modelling**: Using the data given to us as the training data, we have implemented a predictive model, which gives us the production of a particular state or district given its

conditions. This can be particularly beneficial when we can predict the other factors like weather, rainfall, cultivation area, etc using certain techniques and use this data further to predict the production of various crops.

4. **Recommendations to Achieve Higher Production**: If we can find a set of crops that are favourable for a particular area as described earlier, we have summed it up by suggesting methods to achieve higher yields for various crops.

4 Data-Sets And Their Sources

Most of our data is obtained from Kaggle and some government owned websites. We will be using a few data pre-processing techniques as explained later in the Methodology section. The reference for each dataset is given at the end. All the data files used are included in the **"Datasets"** directory in our submission. The following data-sets are used:

1. **Agriculture production[1][2]:** These data-sets contain the District-wise Crop production in India from the year 1997-2014. The files are apy.csv and revenue.csv

2. **Geographical Data[3][4][5]**: These data-sets will contain the Indian district wise data related to Rainfall-patterns[3], Soil nutrient map[4], Mean temperature[5] from the year 2000 to 2014. We have a directory named rainfall that contains all the rainfall data. Other files are Mean_Temp.csv, district_soildata.csv and india87.xls

3. **Prices[8], Minimum support price for Crops[9]** These data-sets contains the State-Wise Prices for various crops. These also include the state-wise Minimum support price different crops from the years 2010-2018. The file are crop_price_per_quintal.csv, prices_all_crops.csv and msp.csv

4. **Area under Production[6] and Production Cost[7]** These data-sets will provide us with crop-wise area under production as well the crop wise production cost for the years 2000-2014. The files are apy.csv and cult_cost.csv, income.csv

5. **Consensus Data[10]** As mentioned earlier, we will be needing the district wise consensus and the literacy data for our analysis. This data-set will provide us with the same for the years 2001 and 2011. The file is literacy.txt.

5 Methodology

5.1 Data Pre-Processing

Data-sets available to us are from various sources and there are many missing values. Also, the units of measurement are different in different datasets and different dataset has different year range. There were some crops whose name were different across different datasets. So, first we needed to do pre-processing such as handling missing values,integrating the datasets in a fixed format. This included techniques such as applying the mean of given years, changing the units and calculation in a better format. The missing values are filled with mean values sometimes, sometimes they are extrapolated or interpolated depending on the requirements. After this data pre-processing, the data was ready to be used for further usage.

5.2 Data Analysis

We have done an extensive study on the crops. We have done both nationwide as well as district wise analysis. We used various data mining concepts for this. We have plotted various graphs for better visualization of the data. For most part we used the line graph, but at some places we have used pie charts, box plots, bar plots as well. We have also use the dotted graph to better visualize the rainfall data.

5.3 Clustering

Firstly, we have tried to do a qualitative visual analysis by visualising through various plots to see correlation between different parameters such as temperature, rainfall, production, cost of cultivation. Then we used some data mining methods to find various peaks and draw inferences from those. Then, we tried to club similar crops and states together so that we can get a better insight of the problem. We have used data mining techniques such as clustering with K-NNs algorithms to see which states, districts, regions or crops are similar in terms of production and demographic factors like soil type, temperature, rainfall, humidity, soil pH, nitrogen levels, etc. The various clusters are depicted in the Analysis and Results section below.

5.4 Predictive Modelling

We have used concepts like Linear regression, decision tree, Random Forest, XGBOOST, ANN to predict crop production based on parameter like cultivation area, rainfall, temperature, season,location etc. As done in all evaluation techniques, data is divided into two parts: Train and Test data. We divided the year-wise data present with us into two parts, the initial years(from years 2002-2010) data is used as the train data and the subsequent data is the Test data. We have run our model on this test set after train and compare the results with the actual data present with us. The efficiency and accuracy of various models are given in the Analysis and Results section.

5.5 Recommendation

As we have already mentioned, we have used the data of year-wise cost of cultivation(per hectare and per quintal). Combined with the crop production data and their price for different crops in each state, we have found which crops are more profitable and which crops price or production are increasing. For instance, a simple evaluation will show that if a crop's cost of cultivation is decreasing but its price is increasing, then that crop is more profitable for farmer to cultivate. Similarly, recommendations can be made by taking into account the weather and the atmosphere of a region, which can be easily retrieved from the data-sets and techniques we are using. We have given the optimal pH ranges, soil types as well as the optimal rainfall for each crops. Apart from this, we have given the state-wise best suited crops given the demographical and economical conditions of the state.

6 Analysis and Results

6.1 Nation-wide Analysis:

6.1.1 Crop Price:

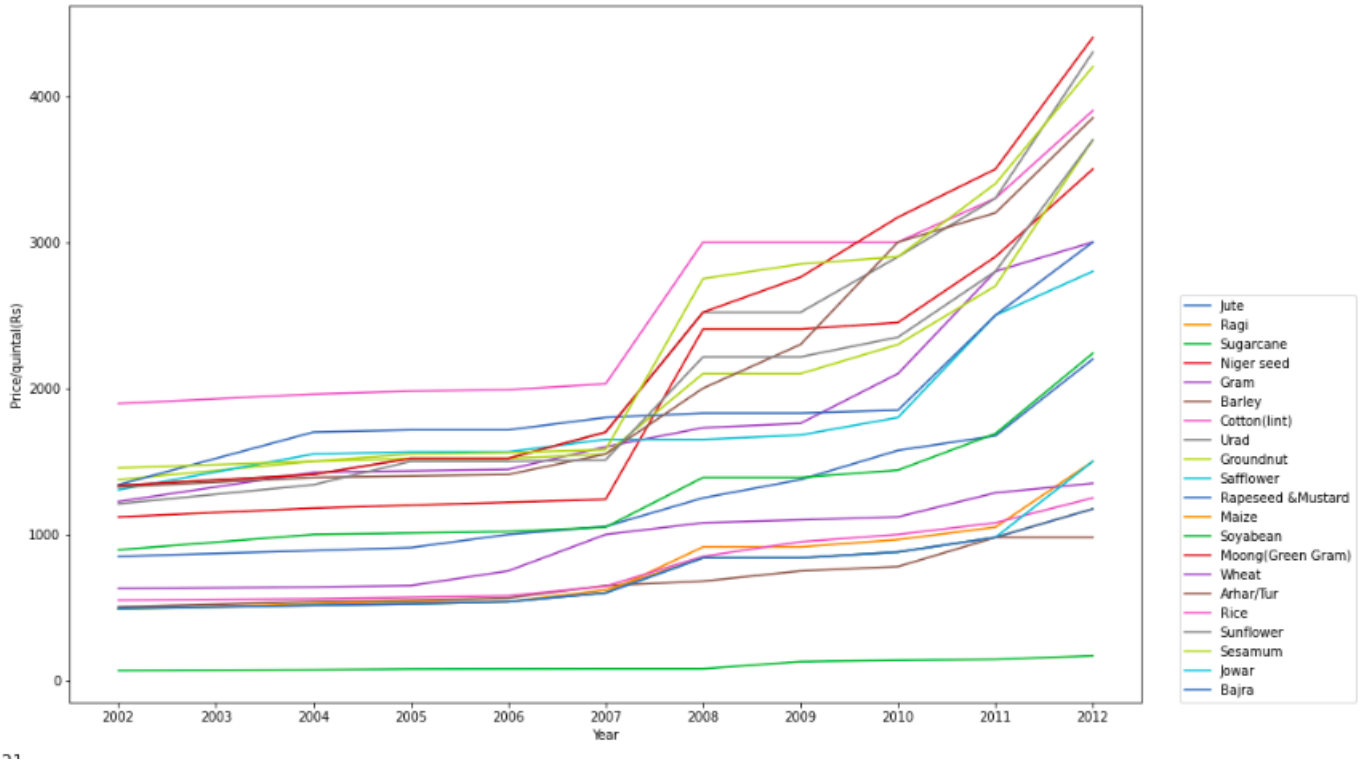


Figure 1: The figure represents the average price per quintal of crops in India, over the years 2002-2012

We can see through the visualization in Figure 1, that there is a sudden increase in prices during the time periods 2007-2008 and 2010-2011. The period 2007-2008 is known as the **World Food Price Crisis**. Root causes for this is known to be the increasing use of bio-fuels in developed countries and an increasing demand for a more varied diet across the expanding middle-class populations of Asia. Similar factors are responsible for increasing cereal prices in 2010-2011. Some, structural in nature, create an imbalance in supply and demand; other are macroeconomic factors related to exchange rates and oil prices.

Also, we can see that certain crops such as sugarcane have very low price as compared to the other crops. This can be attributed to the fact that its yield as well as production is very high, as it will be shown in further sections. On the contrary, certain pulses like Moong and Urad have a really high price. This is because, over the years, as the literacy rate and the awareness has increases, the demand for these crops has increased. People have identified their nutritional value of these crops; therefore the demand; consequently the price of crops has risen. Also, the cost of cultivation of pulses is generally higher(will be shown in later sections) and so are their prices.

6.1.2 Crop Production:

Here(In figure 2), it is clearly evident that sugarcane and cereals make up the majority of share of total production in India(80%). Globally, India is second largest producer of Sugarcane after Brazil. With huge production of sugarcane as well as sugar; largest number of cane farmers; largest consumer of sugar; more than 500 sugar mills and one of the largest sugar exports, India is considered to be a sugar giant. Also, the yield of sugarcane is very high as compared to any other crop. On the other hand, cereals are staple diet of most people in India. This is due to their high nutritive value and good yield as compared to other crops. Therefore, historically, these has come to be one of the most significant crops in India in terms of production.

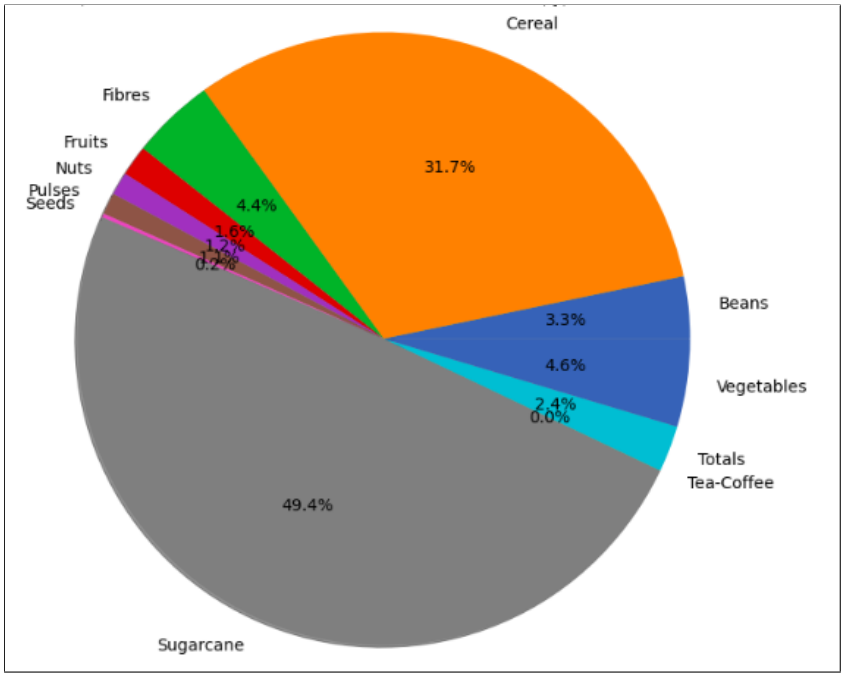


Figure 2: The figure represents the percentage of crops produced in India averaged over years 2002-2012

Also, we can observe through the bar plot(Figure 3) that certain crops such as Sunflower, Niger Seed and Safflower have negative growth rates over the years. All these fall under the category of oil seeds. The absence of assured yield is a reason behind the farmers showing lack of interest in oilseeds. The productivity of oil seeds tend to be on a lower side, therefore India has become more and more dependent on edible oil imports. One of the biggest constraints to raising oilseed ouput has been that the production is largely in rain-fed areas(except for sunflower). Only one fourth of oilseed producing area in India remain under irrigation.

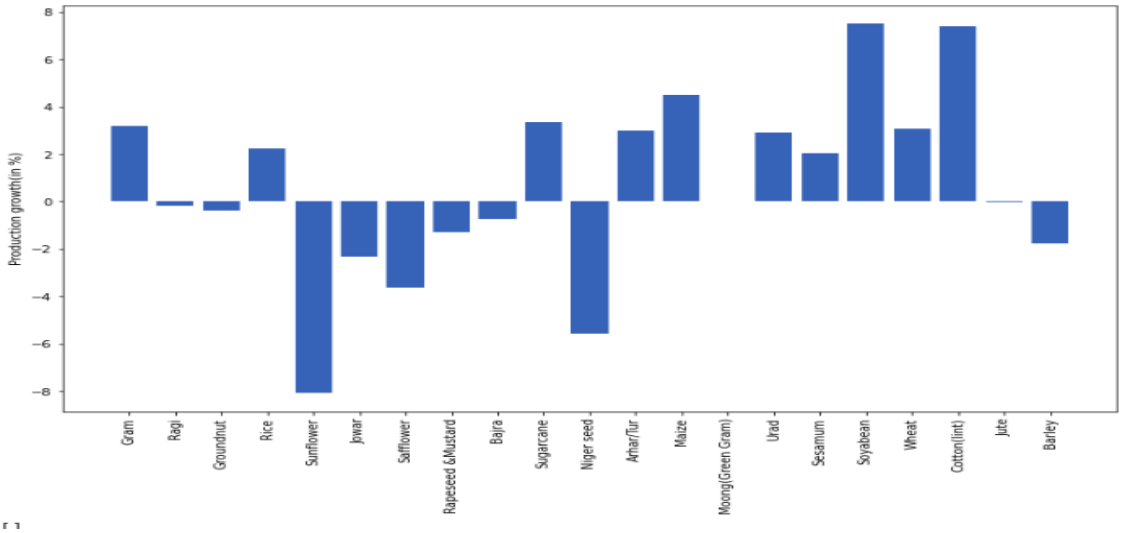


Figure 3: The figure represents the Percentage change in growth of various crops from year 2002-2012

6.1.3 Crop-Wise Cultivation Area:

Cereals cover the most area for production in India, followed by beans. The primary reason for this is that they can be grown in various areas, even in adverse soil and climatic conditions. They also give high yields per acre as compared to other crops. They are not just high in nutritive value, but can be produced with reasonable effort, and can be stored easily in kitchens. Due to this reason, cereals have been one of the most grown crop historically and even now. Similarly, beans have a high protein content, due to which it grown on a wide-scale in India. It also has a lot of industrial usage and are grown in sub-tropical regions and hilly slopes.

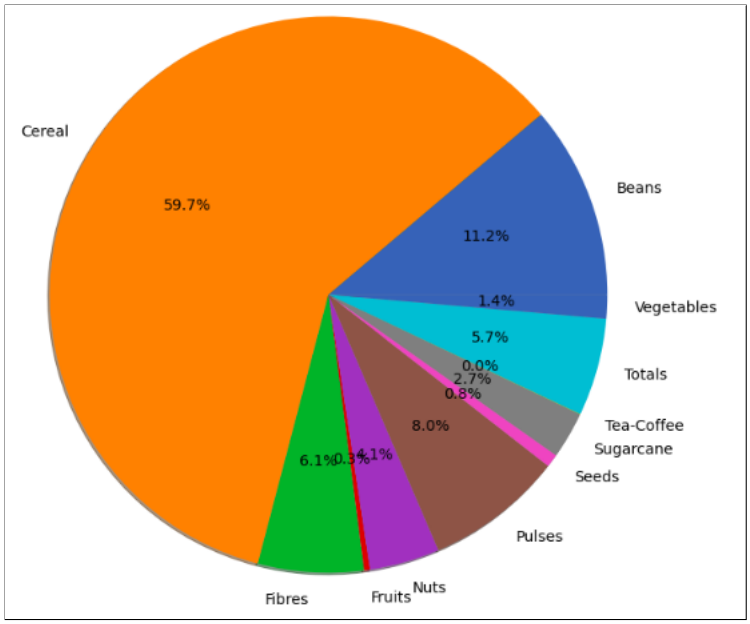


Figure 4: The figure represents the percentage share of various crops in the cultivation area averaged over 2002-2012

6.1.4 Crop Yield:

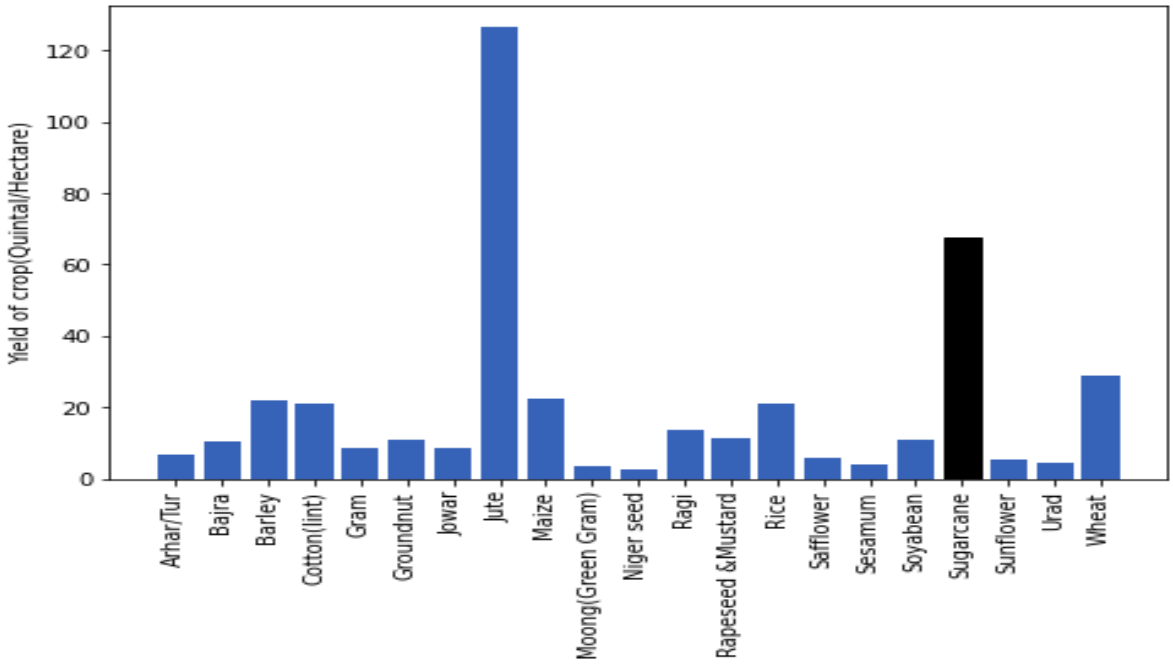


Figure 5: The figure represents the Yield (Production per unit area) for various crops averaged over 2002-2012. Note that yield of crops shown in black bars are divided by ten for better visualization

Sugarcane grows well in humid and hot climate. It can be grown in variety of soils. Therefore, it has a higher yield. The only drawbacks are that it needs manual labour from time of sowing to harvesting and it is soil-exhausting crop and thus needs regular application of manure or fertilizers. Similarly, Jute is also a soil exhausting crop and it grows well in high temperature and heavy rains area such as Wet Bengal. Due to favourable conditions in India and relatively new techniques used because of it being a commercially beneficial crop, the yield of Jute is also very high. As already mentioned above, the yield of pulses and oilseeds is very poor. The reason being that they are water extensive crops and most of the areas in India have inadequate irrigation facilities.

6.1.5 Crop-Wise Revenue:

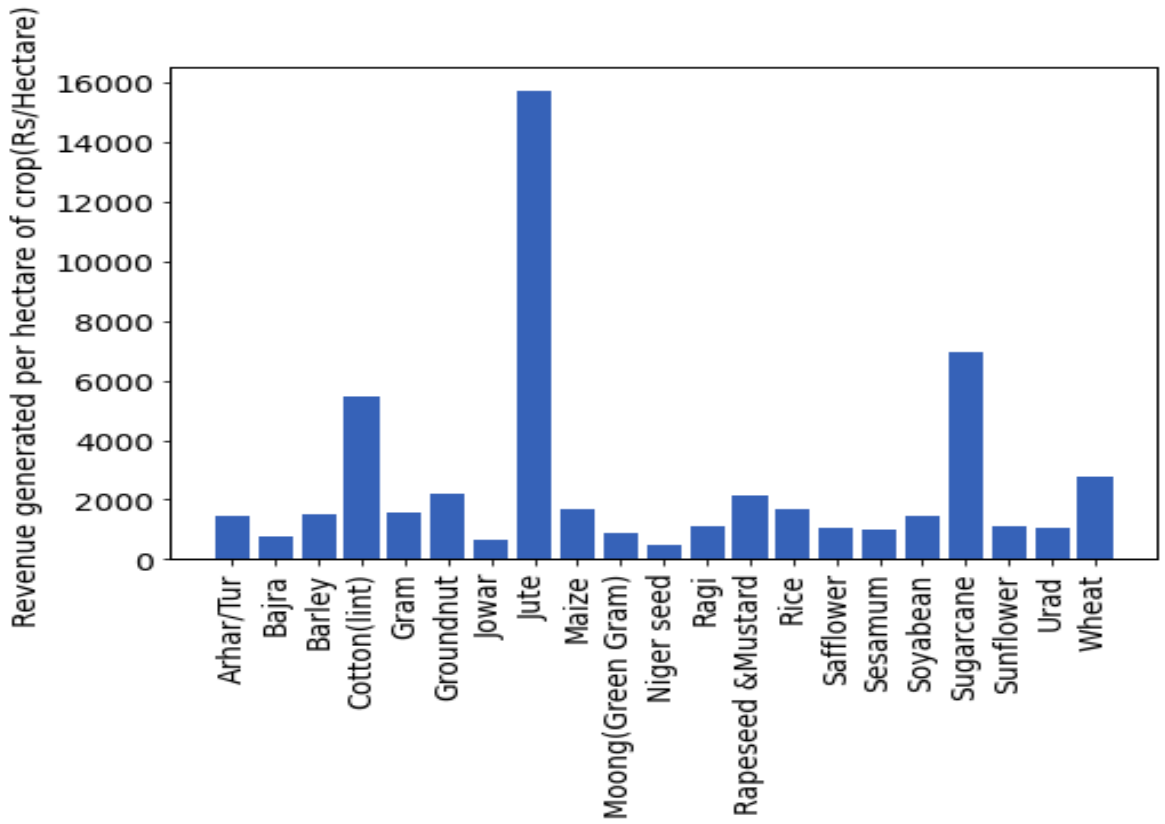


Figure 6: The figure represents the Domestic Revenue per unit area averaged over the years 2002-2012

The revenue shown in the figure is calculated taking into account only the domestic sales and not the exports. For this calculation, the production cost is not taken into account as well. Here, the revenue per hectare of Jute and Sugarcane is again maximum. This can be attributed to their high yield which consequently leads to newer techniques being used in their production. This has lead to them being India's few commercially beneficial crops. India is not only a great consumer of these crops but also one of the main exporters in the world. The revenue generated for Cotton(Lint) is also very high. It is one of the high quality cottons and thus, it is a sold commercially on a large scale at high prices as shown in the Figure 1.

6.1.6 Domestic Earnings:

In figure 7, we have plotted the domestic earnings i.e excluding the export data, for various crops. Here, we have taken into account the cost of production of each crop, year-wise from 2008 to 2012 averaged over all states of India. We have taken the difference between the crop price and the production cost per quintal to calculate the total income per hectare for growing that crop. We can see that income for sugarcane and wheat is high due to their high yield and lower cultivation cost per quintal. On the contrary, pulses have lower yield and thus higher cost of cultivation per quintal and thus lesser earnings per hectare. Therefore, crops such as Sugarcane, Wheat, Maize, Mustard are the most beneficial crops in terms of income.

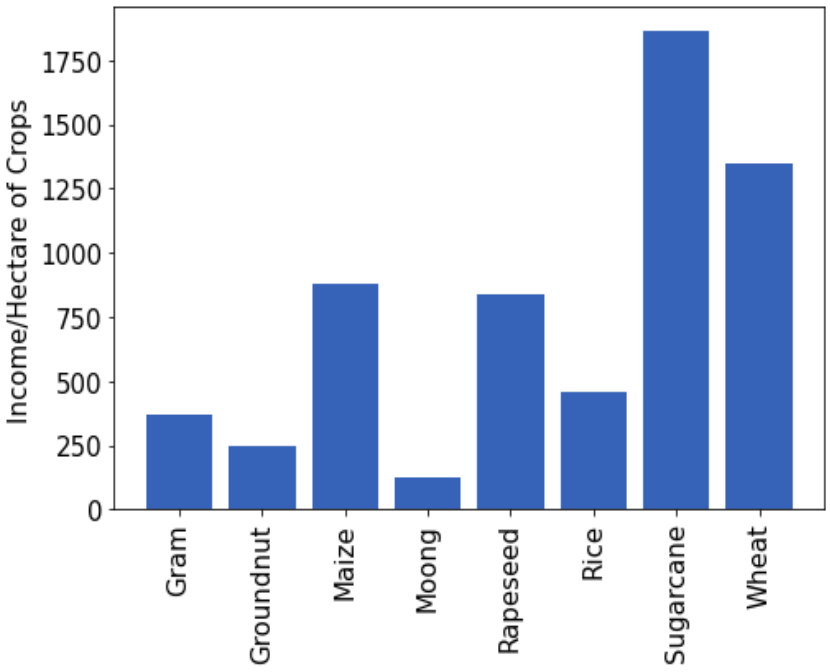


Figure 7: The figure represents the Domestic income per unit area averaged over the years 2008-2012

6.1.7 Cultivation Cost:

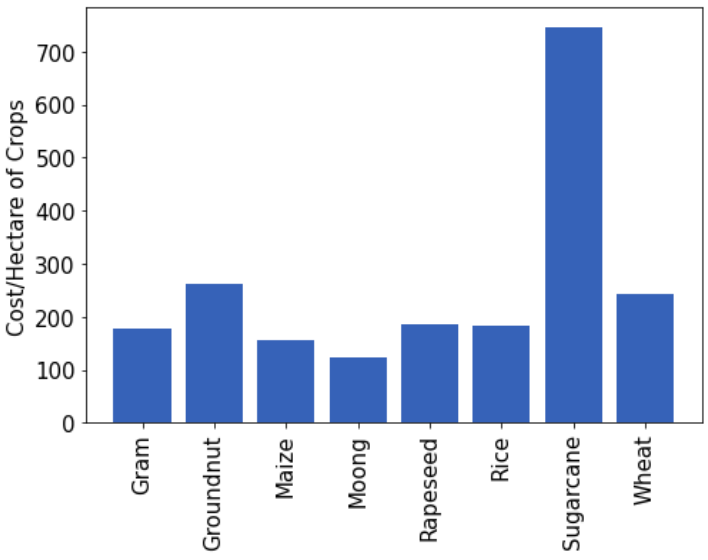


Figure 8: The figure represents the cultivation cost per unit area averaged over the years 2008-2012

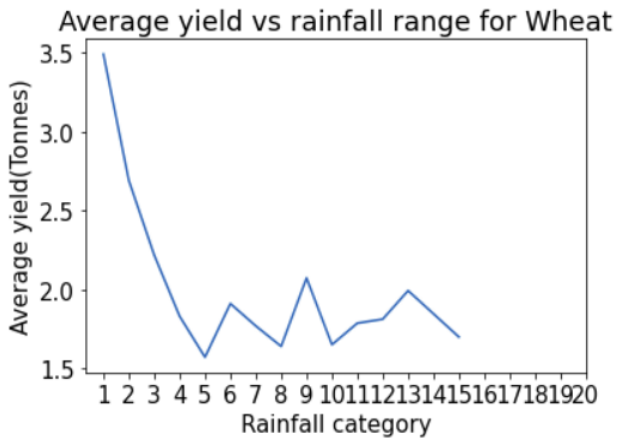
At first glance, the following figure may seem a bit of an outlier. Since sugarcane has been one of the most profitable crop, the cultivation cost was expected to be low. Actually, the cultivation cost of sugarcane is high due to excessive need of fertilizers and it renders the soil useless after that. But the yield(production per area) of sugarcane is so high that it makes up for the high cost. The cultivation cost of other crops is not as high. This is mainly because most of the crop production in India rely on rainfall rather than irrigation. Though it has a certain drawback in the yield, the cultivation costs tend to be on a lower side. Different crops require different amount of rainfall, which is shown in the next section.

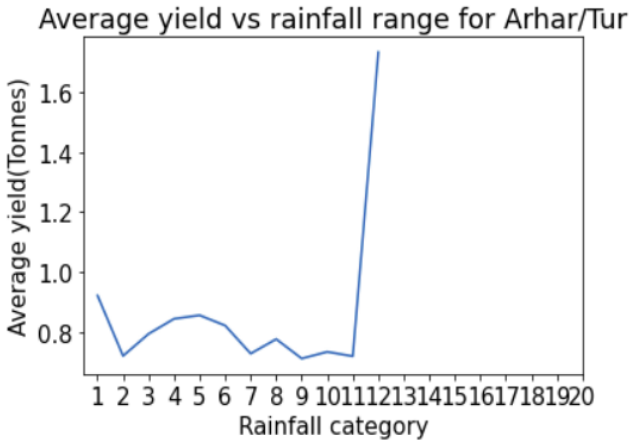
6.1.8 Optimum Rainfall for various crops:

Crop Name	Optimum Rainfall Range(in cm)
Arhar/Tur	3472-3772
Bajra	3472-3772
Barley	181-480
Cotton(lint)	181-480
Gram	181-480
Groundnut	181-480
Jowar	181-480
Jute	3173-3472
Maize	3173-3472
Moong	3173-3472
Niger seed	3173-3472
Ragi	3173-3472
Rice	3173-3472
Safflower	3173-3472
Sesamum	3173-3472
Soyabean	3173-3472
Sugarcane	480-779
Sunflower	480-779
Urad	480-779
Wheat	480-779

Table 1: The following table represents the optimum rainfall category for each crop. The data used was from years 2002-2012

We have used the rainfall data for each district of India across various years and tried to correlate it with the rainfall data for that particular district. We divided the amount of rainfall into 20 different categories and found which category has the highest yield for that particular crop. We tried to find the optimal rainfall for the crop, in which the yield of that particular crop is maximum. For this, we have ignored other factors such as temperature and soil pH for the time being. There are some water extensive crops such as Arhar, Bajra. They are generally grown in areas with high rainfall. On the other hand, crops such as Wheat, Urad, Sugarcane can be grown in areas with lower rainfall. We have shown the yield v/s rainfall plots for a couple of crops below for a better visualization.



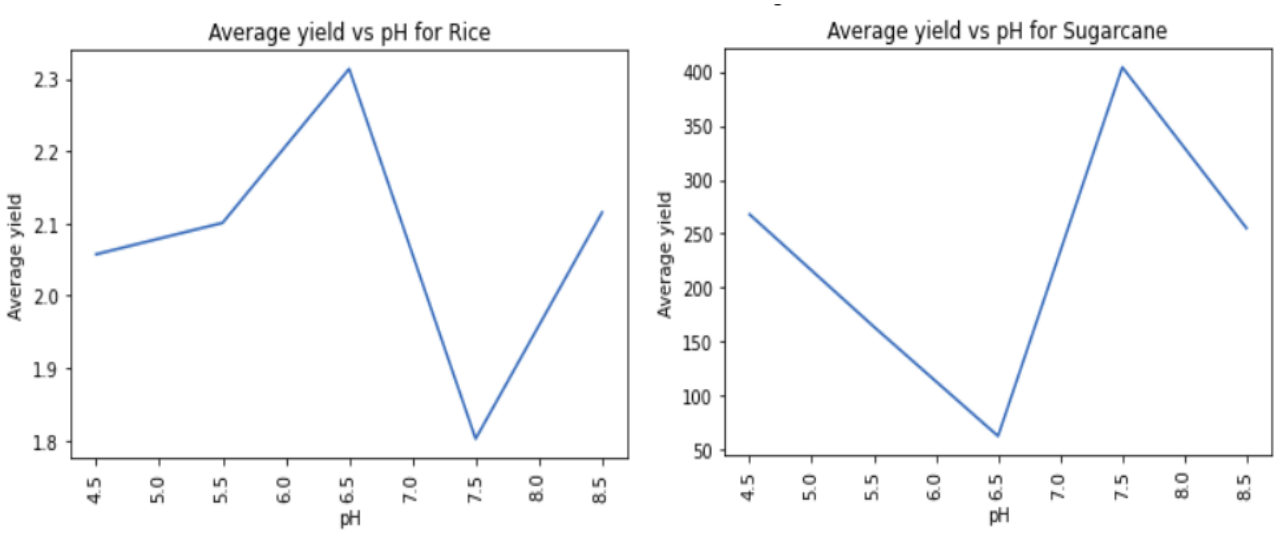


6.1.9 Optimum Soils for various crops

Here, we have used the soil type and soil pH data for each district of India across various years and tried to find the best soil to grow crops in and the optimal pH values for each of these major crops. We had in total 21 different soil types. The data for the production of every crop across each district is correlated with the soil type and the pH values of the soil found in that region. For this, we have ignored the other factors like temperature, rainfall, cultivation costs, etc and tried to do the analysis only on the basis of soil. As we can see, most the crops tend to grow(have the best yield) in regions from pH of 5 to 9. Too acidic or too basic soil tend to hinder the water intake capacity of the crops by interfering with the natural osmosis of crops. Due to this, most crops are best grown in near to neutral pH, i.e seven.

Crop Name	Optimum Soil Types	Optimum pH
Arhar/Tur	Calcerous & Shallow Black	8-9
Bajra	Coastal Alluvial & Deltaic Alluvium & Alluvial River	8-9
Barley	Gray Brown & Saline and Alkaline & Alluvial River	8-9
Cotton(lint)	Deltaic Alluvium & Red and Gravely & Coastal Alluvial	8-9
Gram	Black (Karail) & Deltaic Alluvium & Coastal Alluvial	8-9
Groundnut	Desert & Deltaic Alluvium & Coastal Alluvial	8-9
Jowar	Black (Karail) & Coastal Alluvial & Medium Black	8-9
Jute	Red and Gravely & Red & Tarai	5-6
Maize	Black (Karail) & Coastal Alluvial	4-5
Niger seed	Tarai & Coastal Alluvial	8-9
Ragi	Black (Karail) & Saline and Alkaline & Alluvial River	6-7
Rice	Black (Karail) & Coastal Alluvial & Tarai	6-7
Safflower	Tarai & Coastal Alluvial & Saline and Deltaic	8-9
Sesamum	Deltaic Alluvium & Coastal Alluvial & Calcerous	7-8
Soyabean	Red and Gravely & Black (Karail) & Deep Black	4-5
Sugarcane	Alluvial River & Black (Karail) & Coastal Alluvial	7-8
Sunflower	Gray Brown & Saline and Alkaline & Calcerous	8-9
Urad	Calcerous & Deltaic Alluvium	8-9
Wheat	Deltaic Alluvium & Saline and Alkaline & Alluvial River	8-9

Table 2: The following table represents the optimum soil type and pH values for each crop. The data used was from years 2002-2012



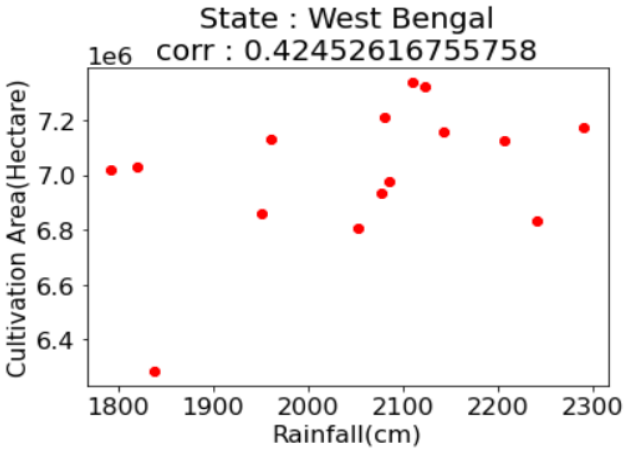
6.2 Geographical Analysis:

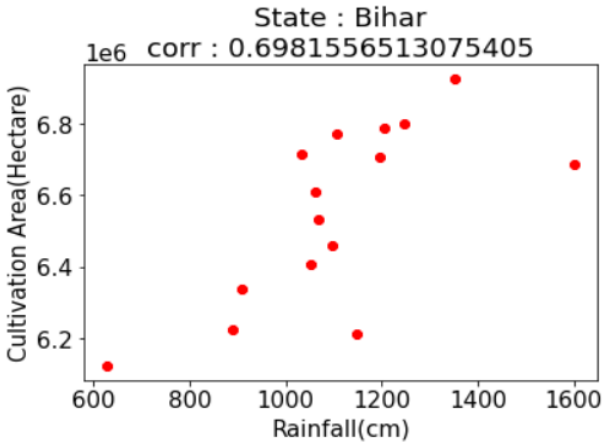
6.2.1 Rainfall and Cultivation Area

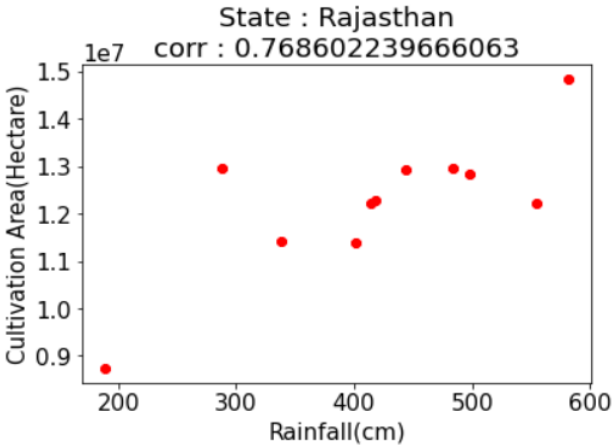
As already discussed above, most of the cultivation in India depends on the rainfall. Most of the small scale farmers are dependent on the rainfall for the growth of the crops. Also, we have seen in the Table 1, how most of the crops are water intensive, i.e they have medium to high water requirements. Therefore, the area under cultivation across states should have some correlation with the rainfall.
We tried to find this correlation taking the rainfall data of all the districts over many years and analyzing their cultivation area in total. The correlation coefficient has a range from 0 to 1, where 1 being the most correlation and vice versa. We could categorize the states into two major types, one which have sufficient rainfall spread across the whole state and the other which had inadequate rainfall in most of the districts of the states. We could infer two things from that:

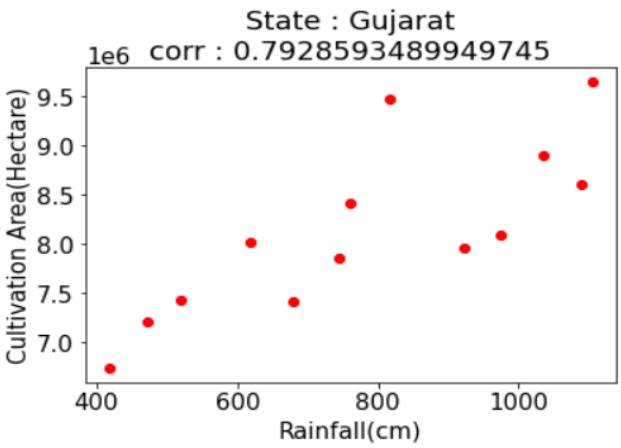
- For the high rainfall states such as West Bengal, Arunachal Pradesh, Odisha, the correlation coefficient was lower. This is because most of the districts of these states have adequate rainfall and therefore they can fulfill the basic requirements of water for most of the crops. Due to this, the cultivation can be done in practically every district, as far as rainfall is concerned.

- For the low rainfall states such as Rajasthan, Gujarat, the correlation coefficient was comparatively higher. This is because only some of the districts fulfill the basic requirements of water for most of the crops and therefore, it is preferred to grow crops in those districts.



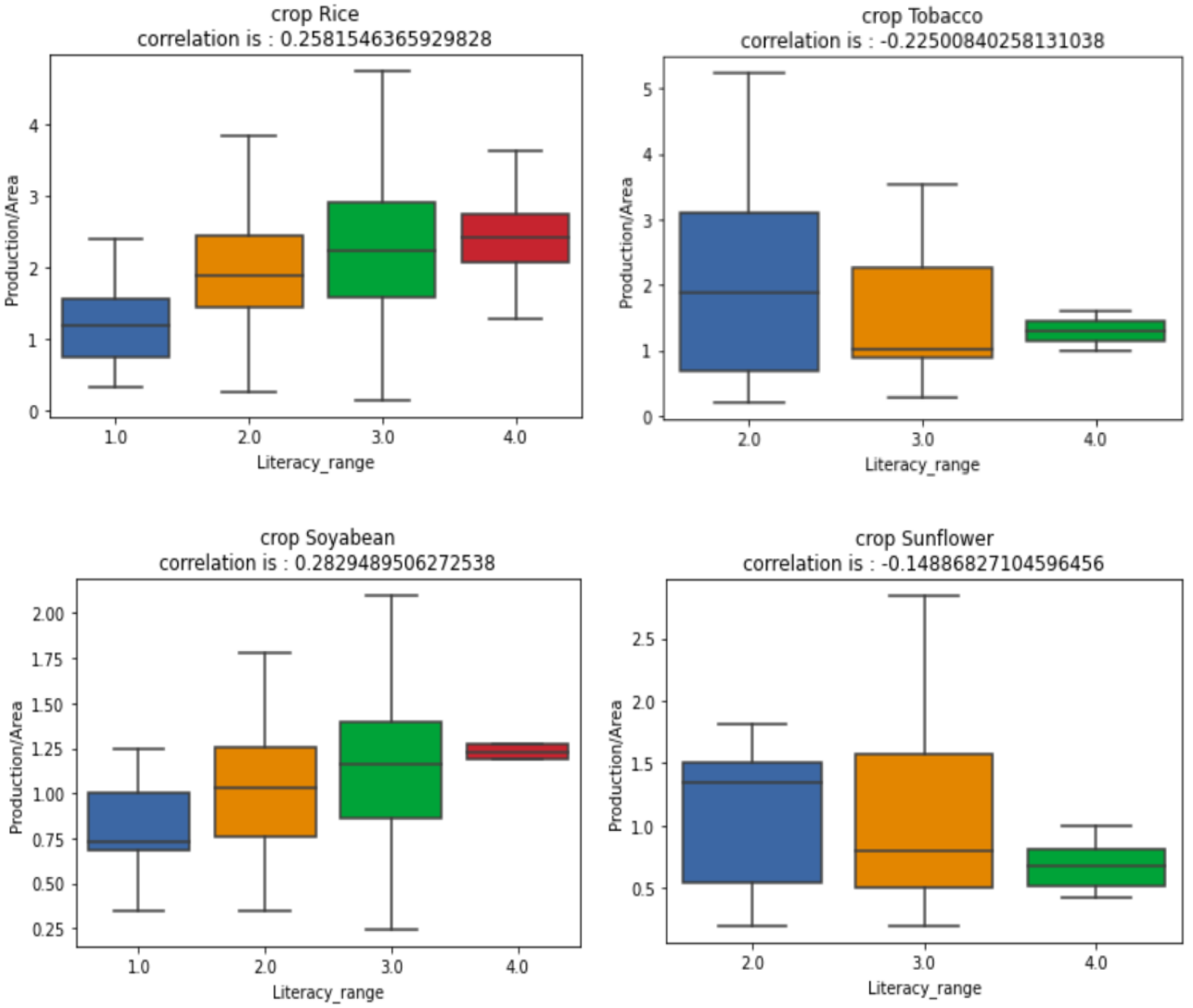






6.2.2 Correlation of Literacy Rate with Production

It might not be intuitive that the literacy rate of an area can have any correlation with the production of crops in that area. Surprisingly enough, some crops did have a good correlation with the literacy rates of the area. We have used a correlation factor to quantify it. High correlation means the yield(Production/Area) increase with literacy rate and negative correlation means the yield decreases. We have used box plots to visualize them. We have divided the literacy rate into 4 ranges: 0-50%, 50-70%, 70-85% and above 85%. Upper and lower edge of the box represents the 25 and the 75 percentile in that particular range and the middle line represents the median value.
We see that crops such as soyabean and rice have a high(and positive) correlation with the literacy rate. This can be attributed to the fact that they are pretty healthy crops and the farmers tend to find a strong market in the nearby areas because of people being aware of their benefits. On the contrary, crops such as tobacco and sunflower(oilseeds) tend to have more market among the people with lower literacy rates. Please note that we tried to correlate it with the literacy rate of the district and not of the farmers only, therefore more literacy rate need not necessarily mean they would be technologically advanced.



6.2.3 Crops Growing in a Similar Environment

We tried to cluster the similar crops(In figure 13) using t-SNE visualization on the basis of the environmental conditions in which the yield of crops is better. We have used KNN-clasifier for the same. The features or the factors we used are Season(Kharif, Rabi, Summer), type of soil, soil pH, rainfall and temperature. As shown in figure 13, we got 8 clusters based on this. The number 8 is chosen based on the Silhouette Coefficient values as shown in the figure 14. Silhouette score is metric which is used to calculate the goodness of the clustering technique. The values range from -1 to 1. Higher the value of Silhouette score, better is the cluster. As we can see in Figure 14, the value is maximum for the number of clusters equal to eight. Each cluster shows one major characteristics(major feature) upon which that particular cluster is based. The clusters and their characteristics are:

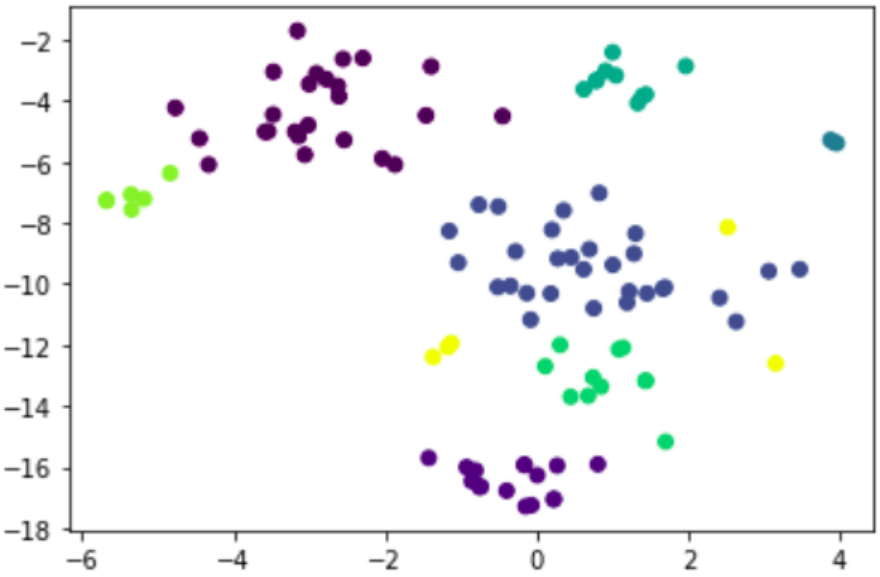


Figure 13: The figure represents the clusters of similar crops, i.e crops that can be grown in similar environmental conditions

- **Cluster 1:** These are the crops which are grown in Kharif season. Alluvium river and red soils are the optimum soil types for these crops and occurs at places with rainfall range from 1000-1500mm. The main crops are **Groundnut, Horse-gram, Jute, Maize, Moong(Green Gram), Ragi, Rice, Sesamum, Sunflower and Urad.**

- **Cluster 2:** These are the crops which are grown in both Rabi and Kharif season. Coastal Alluvium and red soils are the optimum soil types for these crops and occurs at places with rainfall range from 300-700mm. The main crops are **Ash Gourd, Beet Root, Ber, Bitter Gourd, Carrot, Drum Stick, Jack Fruit, Lab-Lab, Litchi, Pear, Pineapple, Pump Kin, Redish, Ribed Guard, Snak Guard and Yam.**

- **Cluster 3:** These are the crops which are grown throughout the year. These crops grow in majorly in Alluvium river soil. Other soil types are red, black(Karail), coastal alluvial, deep black. Optimal rainfall is range from 900-1000mm. The main crops are **Banana, Beans & Mutter(Vegetable), Bhindi, Bottle Gourd, Brinjal, Cabbage, Cauliflower, Citrus Fruit, Coriander, Cucumber, Dry chillies, Dry ginger, Garlic, Ginger, Grapes, Guar seed, Mango, Oilseeds total, Onion, Orange, Other Citrus Fruit, Other Fresh Fruits, Other Vegetables, Papaya, Peas (vegetable), Pome Fruit, Potato, Sannhamp, Sugarcane, Sweet potato, Tobacco, Tomato, Turmeric, Water Melon and other fibres**

- **Cluster 4:** These are the crops which are grown in mostly Kharif season. These crops grow in majorly in Alluvium river soil. Other soil types are red an black. Optimal rainfall is range from 900-1300mm. The main crops are **Arhar/Tur, Bajra, Castor seed, Cotton(lint),**

Cowpea(Lobia), Jowar, Mesta, Moth, Niger seed, Other Cereals & Millets, Other Kharif pulses, Small millets, Soyabean, other misc. pulses and other oilseeds.

- **Cluster 5:** These are the crops which are grown throughout the year. These crops grow in majorly in Alluvium type soil. These crops require heavy rainfall requirements. Optimal rainfall is range from 1300-2700mm. The main crops are **Arcanut (Processed), Arecanut, Atcanut (Raw), Black pepper, Cardamom, Cashewnut, Cashewnut Processed, Cashewnut Raw, Coconut , Rubber, Tapioca**
- **Cluster 6:** These are the crops which are grown in Kharif season. Calcerous and Tarai soils are the optimum soil types for these crops and occurs at places with rainfall around 1200mm. The main crops are **Lentil, Pulses and Total foodgrain.**
- **Cluster 7:** These are the crops which are grown in Rabi season. Black(Deep) and Alluvial soils are the optimum soil types for these crops and occurs at places with rainfall range from 800-1000mm. The main crops are **Barley, Gram, Khesari, Linseed, Masoor, Other Rabi pulses, Peas & beans (Pulses), Rapeseed &Mustard, Safflower and Wheat.**
- **Cluster 8:** These are the crops which are grown in Kharif season. Black(Deep and Medium) and Alluvial river soils are the optimum soil types for these crops and occurs at places with rainfall range from 700-1100mm. The main crops are **Korra, Lemon, Pome Granet, Samai, Sapota and Varagu.**

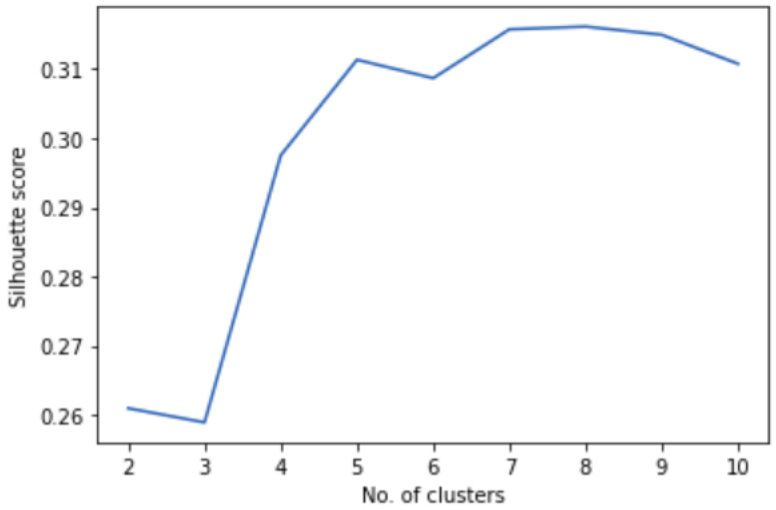
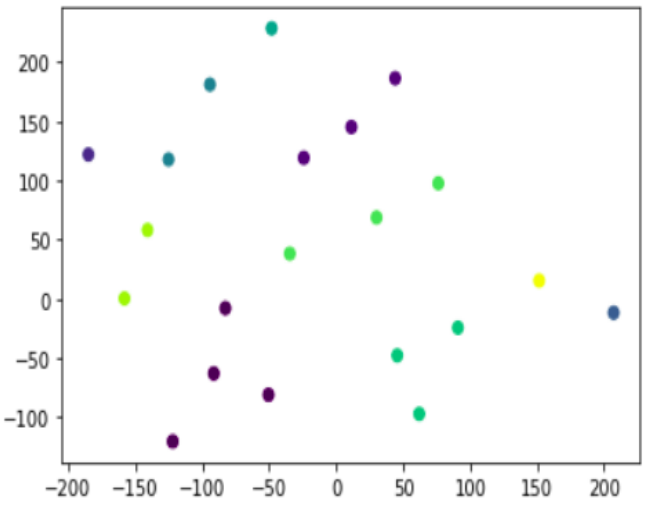


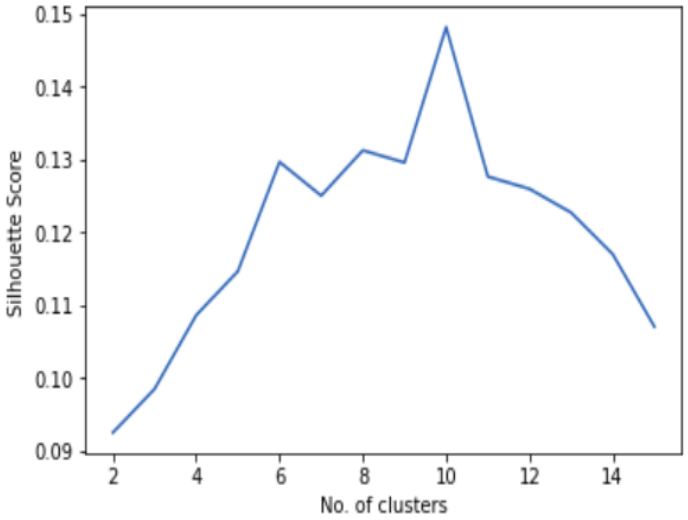
Figure 14: The figure represents the Silhouette score v/s the number of clusters.

6.2.4 Similar States And Their Production

Similar to what is done above, we tried to cluster the similar states(as shown in Figure 15) using t-SNE visualization. The features used were same as above, i.e soil type share, rainfall, pH range. We have used KNN-clasifier, as shown above. As shown in figure 15, we got 10 clusters based on this. The number 10 is again chosen based on the Silhouette Coefficient value(as shown below). Each cluster shows some major characteristics(major feature) upon which that particular cluster is based. The clusters and their characteristics are:



(a) Figure 15: Cluster of similar districts



(b) Figure 15(b): Silhouette Value v/s No. of Clusters

- **Cluster 1:** The states are Haryana, Punjab, Rajasthan and Uttar Pradesh. District in these states have annual rainfall in the range 190-850 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 2:** The states are Jharkhand, Odisha and Tamil Nadu. District in these states have annual rainfall in the range 850-1500 mm and the major soil is red with pH range of 4-6.
- **Cluster 3:** The states is Maharashtra. District in these states have annual rainfall in the range 520-1800 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 4:** The state is Kerala. District in these states have annual rainfall in the range 2170-3500 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 5:** The states are Bihar and Telangana. District in these states have annual rainfall in the range 520-1500 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 6:** The state is Uttarakhand. District in these states have annual rainfall in the range 1180-1840 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 7:** The states are Karnataka, Tripura and West Bengal. District in these states have annual rainfall in the range 1180-2500 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 8:** The states are Chhattisgarh, Himachal Pradesh and Madhya Pradesh. District in these states have annual rainfall in the range 850-1500 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 9:** The states are Andhra Pradesh and Gujarat. District in these states have annual rainfall in the range 520-1120 mm and the major soil is alluvial with pH range of 6-8.
- **Cluster 10:** The state is Arunachal Pradesh. District in these states have annual rainfall in the range 2170-3160 mm and the major soil is alluvial with pH range of 6-8.

6.3 Prediction

We have used various prediction models to predict the production of various crops given the environmental conditions of district. Since, there are models to estimate the environmental conditions of various, our model can be used in conjunction with those models to predict the production of various crops for a particular district. The methodology has already been explained in the section 5.4 above, we will just show the results here. The factors taken into account are the mean temperature, the cultivation area, the soil type, soil pH, rainfall, the season the crop. The models used are the Random Forest, XgBoost, Decision Tree and Linear Regression. The data from 2002 to 2010 was taken as the train data and the rest data was taken as the test data. We had varied accuracy, i.e the values of R-square and Mean Absolute Error and we chose the one with least error i.e the Random Forest Model. All the models and their accuracy are depicted as follows:

6.3.1 Random Forest

This is the current model which we are using for the prediction. This is a Random Forest Model and its R-squared value came out be 0.94 and Mean Absolute Error was 24796. This model has the highest accuracy and lowest error of all the four models we trained, therefore this model was chosen.

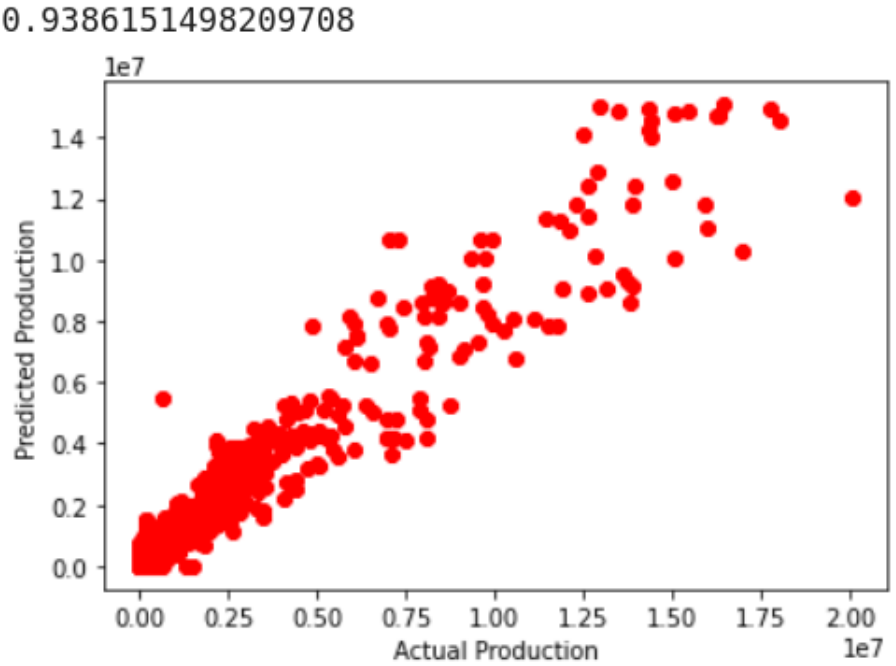


Figure 16: Prediction using random forest

6.3.2 XGBoost

This is the second best model that we trained. The model is an XGBoost Model and its R-squared value came out be 0.93. The Mean Absolute Error was calculated as 31355.

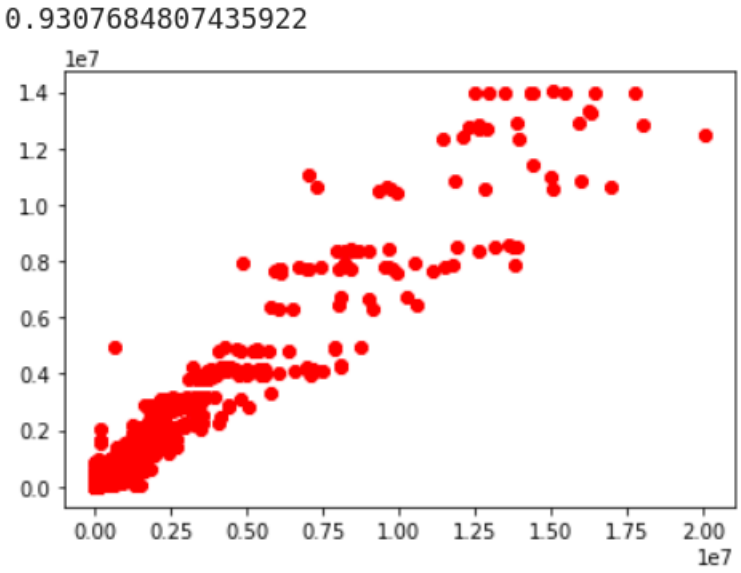


Figure 17: Prediction using XgBoost

6.3.3 Decision Tree

The Decision Tree Model did pretty good results as well. The R-squared value of the model came out to be 0.91 which signifies the model has indeed a very high accuracy. The Mean Absolute error was found to be 31189.

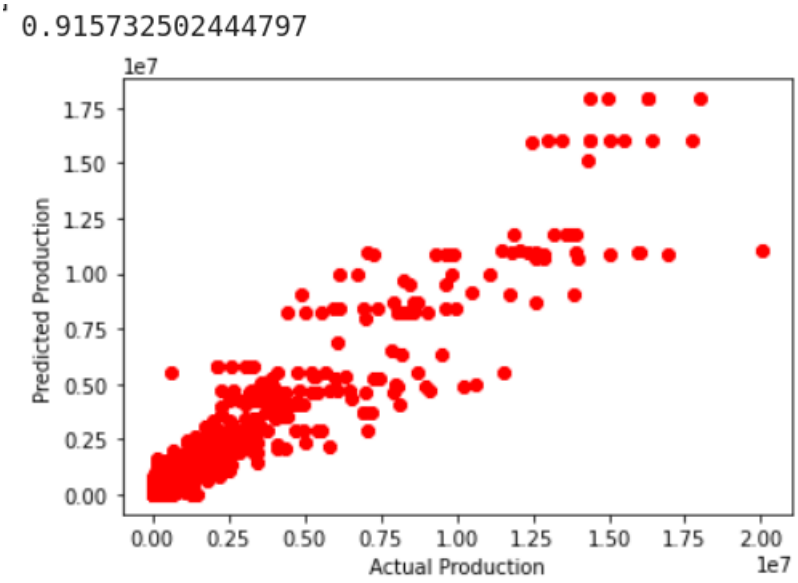


Figure 18: Prediction using Decision tree

6.3.4 Linear Regression

This is the least accurate model that we trained. The prediction was pretty bad and thus, we rejected this model out-rightly. The R-squared value of the model came out to be 0.18 and Mean Absolute Error of 110880 which signifies that model was inaccurate.

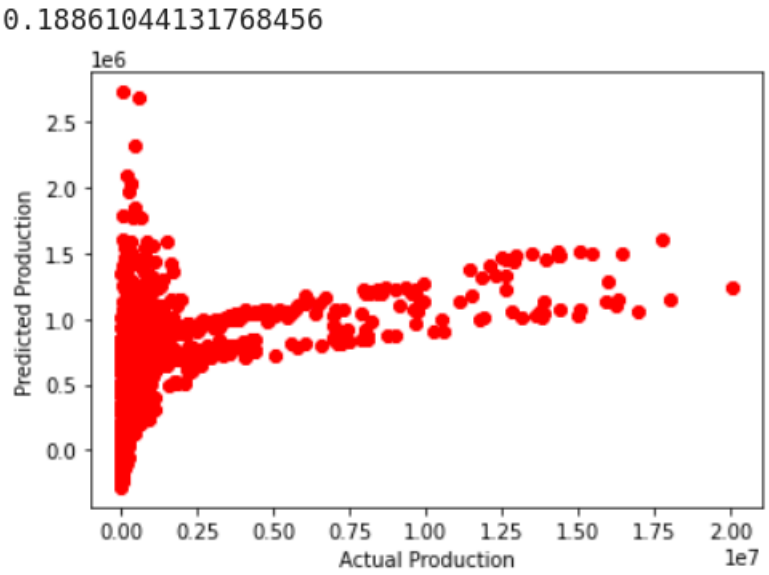


Figure 19: Prediction using Linear Regression

6.4 Recommendation

State	Crop1	Crop2	Crop3	Crop4	Crop5	Crop6	Crop7	Crop8	Crop9	Crop10	Crop11	Crop12
Haryana	Sugarcane	Sweet potato	Turmeric	Banana	Wheat	Cotton(lint)	Onion	Garlic	Potato	Barley	Other Vegetables	Rice
Punjab	Sugarcane	Sweet potato	Turmeric	Banana	Wheat	Cotton(lint)	Onion	Garlic	Potato	Barley	Other Vegetables	Rice
Rajasthan	Sugarcane	Sweet potato	Turmeric	Banana	Wheat	Cotton(lint)	Onion	Garlic	Potato	Barley	Other Vegetables	Rice
Uttar Pradesh	Sugarcane	Sweet potato	Turmeric	Banana	Wheat	Cotton(lint)	Onion	Garlic	Potato	Barley	Other Vegetables	Rice
Jharkhand	Tapioca	Cabbage	Onion	Arhar/Tur	Potato	Maize	Sugarcane	Sweet potato	Banana	Coconut	Jute	
Odisha	Tapioca	Cabbage	Onion	Arhar/Tur	Potato	Maize	Sugarcane	Sweet potato	Banana	Coconut	Jute	
Tamil Nadu	Tapioca	Cabbage	Onion	Arhar/Tur	Potato	Maize	Sugarcane	Sweet potato	Banana	Coconut	Jute	
Maharashtra	Tomato	Maize	Sugarcane	Banana	Mango	Wheat	Cotton(lint)	Onion	Grapes	Rice		
Kerala	Sugarcane	Sweet potato	Banana	Pineapple	Mango	Tapioca	Coconut	Papaya	Garlic	Potato		
Bihar	Papaya	Onion	Sugarcane	Sweet potato	Banana	Jute	Potato	Grapes	Coconut			
Telangana	Papaya	Onion	Sugarcane	Sweet potato	Banana	Jute	Potato	Grapes	Coconut			
Uttarakhand	Sugarcane	Turmeric	Ginger	Wheat	Onion	Dry ginger	Tobacco	Potato	Total foodgrain	Rice		
Karnataka	Sugarcane	Banana	Citrus Fruit	Grapes	Coconut	Dry ginger	Jute	Potato	Mesta	Jute	mesta	
Tripura	Sugarcane	Banana	Citrus Fruit	Grapes	Coconut	Dry ginger	Jute	Potato	Mesta	Jute	mesta	
West Bengal	Sugarcane	Banana	Citrus Fruit	Grapes	Coconut	Dry ginger	Jute	Potato	Mesta	Jute	mesta	
Chhattisgarh	Sugarcane	Sweet potato	Banana	Ginger	Papaya	Onion	Dry ginger	Potato				
Himachal Pradesh	Sugarcane	Sweet potato	Banana	Ginger	Papaya	Onion	Dry ginger	Potato				
Madhya Pradesh	Sugarcane	Sweet potato	Banana	Ginger	Papaya	Onion	Dry ginger	Potato				
Andhra Pradesh	Wheat	Papaya	Onion	Garlic	Potato	Sugarcane	Banana	Pome Fruit	Coconut	Grapes		
Gujarat	Wheat	Papaya	Onion	Garlic	Potato	Sugarcane	Banana	Pome Fruit	Coconut	Grapes		
Arunachal Pradesh	Maize	Sugarcane	Turmeric	Soyabean	Wheat	Dry chillies	Pulses total	Dry ginger	Potato	Rice		

Here, we have made recommendation on what would be the suitable crops to grow in each state. This is created by taking into account the suitable conditions and consequently the yield for the crops. The suitable conditions include the rainfall, soil types, soil pH, cultivation area, the price and the temperature of that particular state. As we can see that Sugarcane is the most preferred crop in most of the states. We have already explained the reasons for this.

7 Discussion

Agriculture sector contributes a significant share in the country's economy and hence it becomes very important to device strategies and planning to get the maximum potential production with the given conditions. This requires using a proper mechanism to understand large datasets of hundreds of districts and states pertaining to so many years. To simplify the process, we break down this into a set of variables like rainfall amount, soil type, soil pH, area of cultivation, production size, etc. To get a primitive idea about the trends and patterns in the data we start with plotting basic plots between many of such correlated variables. This helps us understand if or how the change in our variables affect our observation of others. Moreover, the structure of these basic plots gives us a gateway to visualise the data and guide us for the direction to dig deeper in the data. After understanding and visualising the co-dependency among different variables in the data, we need to exploit these patterns to predict various dependent variables in any given future scenarios. This can give us not only a look-ahead for taking various decisions like crop selection, deciding cultivation area, etc. but also get a picture of the amount of crop production beforehand. We used different ML models like Linear Regression, Random Forest, XgBoost and Decision Tree, to generate the prediction model by observing which model fits our data the best and gives the most accurate test results.

8 Future Direction

In this project, we attempted to analyze every different aspect of the data associated with the problem we chose, limited by the tools we used. Despite this, there still remains a scope of improvement to gain an even more deeper and accurate understanding of the farming production cycle. Firstly, when deriving plots and clusters based on the cost price of production and MSP of different crops, we did not take into account the inflation rate inherently incorporated into the data. Doing this normalisation will give us a more standard ruler to compare various plots over any given temporal space. To do this correction, we would require the data of inflation rate in India over the given set of years, and then use this to scale up (or down) the price/MSP data. Secondly, we can do an improvement in the the way we handle the literacy data. Currently, we are using the literacy rate of the local areas to understand its influence on the farming production and output. This can be improved further by studying the education and literacy levels specifically among the farming community as they are directly associated with the agriculture industry and will affect the production patterns owing to a better access to technological tools. Thirdly, we can employ the fertilizers utilisation dataset to learn better crop recommendations, because as we know, the type of fertilisers used significantly affects the soil condition and consequently the choice of crops to be cultivated. We have not used the above dataset in our project because of unavailability of any such dataset which is compatible with the other datasets that we have used. For similar reasons, we have also not been able to include datasets having information about winds and humidity. All these further integrations can help us upgrade our crops recommendation model and make it as precise and accurate as possible.

References

[1] https://www.kaggle.com/abhiseklewan/crop-production-statistics-from-1997-in-india?select=apy.csv

[2] https://www.kaggle.com/srinivas1/agricuture-crops-production-in-india?select=produce.csv

[3] https://www.kaggle.com/rajanand/rainfall-in-india?select=district+wise+rainfall+normal.csv

[4] http://www.mospi.gov.in/sites/default/files/reports_and_publication/statistical_publication/EnviStats/b14_Chapter%202.pdf

[5] https://data.gov.in/resources/monthly-seasonal-and-annual-mean-temp-series-1901-2017

[6] https://data.gov.in/catalog/area-under-principal-crops-all-india-and-state-wise?filters%5Bfield_catalog_reference%5D=85951 & format=json & offset=0 & limit=6 & sort%5Bcreated %5D=desc

[7] https://www.indiastat.com/agriculture-data/2/cost-of-cultivation-production/32320/stats.aspx

[8] https://agmarknet.gov.in/PriceTrends/SA_Pri_Month.aspx

[9] https://data.gov.in/catalog/minimum-support-priceprocurement-price-crops-crop-year-basis?filters%5Bfield_catalog_reference%5D=89010 & format=json & offset=0 & limit=6 & sort%5Bcreated%5D=desc

[10] https://en.wikipedia.org/wiki/List_of_Indian_states_and_union_territories_by_literacy_rate