A Project on

CROP PRODUCTION PREDICTION

Submitted for partial fulfilment of award of

MASTER OF SCIENCE

DR. HOMI BHABHA STATE UNIVERSITY, MUMBAI

Degree in

DATA SCIENCE

By

Simran Kishan Kanojia

Under the Guidance of

Ms. Prachee Angane

Department of Mathematics

The Institute of Science

Mumbai – 400032

May, 2023

Declaration

I hereby declare that the project work entitled "**Crop Production Prediction**" carried out at the Department of Mathematics, The Institute of Science, Mumbai, is a record of an original work done by me under the guidance of Dr. Selby Jose, The Institute of Science, and this project work is submitted in the partial fulfilment of the requirements for the award of the degree of Master of Science in Data Science, Dr. Homi Bhabha State University. The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

SIMRAN KANOJIA

DS2216

Acknowledgment

One thing that deserves my utmost gratitude is the opportunity to work on this research project.

Above all, I would like to extend my heartfelt gratitude to my parents for their unwavering support and encouragement throughout. Their unique insights and guidance were priceless in our success of this project, and I feel incredibly lucky to have had such a wonderful support system.

I am immensely grateful to **Dr. Selby Jose**, Head of the Department of Mathematics, and the entire Department of Mathematics for providing us with this unique opportunity to work on this project.

I would like to take this opportunity to express my heartfelt gratitude to our project guides, **Ms. Prachee Angane** and **Ms. Juilie Shinde** for their invaluable contributions to this project. Their expertise in the field has been invaluable in helping me to navigate the challenges and complexities of this project, and their guidance and feedback have helped me to refine my ideas and approach to the work. They have been a source of inspiration and encouragement, and I am truly grateful for their unwavering commitment to my success. I would like to extend my deepest appreciation to everyone who came across and motivated and supported. I am truly grateful for the opportunity to work with the talented and dedicated group members. Their hard work, dedication, and commitment to excellence have been a constant source of inspiration and motivation for me.

This project as a whole has been a journey of growth and learning for me, and I am grateful for the experience. Finally, I would like to thank my mentor for their guidance and support, which helped me navigate the challenges and opportunities that arose during the project.

SIMRAN KISHAN KANOJIA

Contents

Chapter 1

Introduction

Agriculture, since its invention and inception, be the prime and pre-eminent activity of every culture and civilization throughout the history of mankind. It is not only an enormous aspect of the growing economy, but it's essential for us to survive. It's also a crucial sector for Indian economy and also human future. It also contributes an outsized portion of employment. Because the time passes the requirement for production has been increased exponentially. So as to produce in mass quantity people are using technology in an exceedingly wrong way. New sorts of hybrid varieties are produced day by day. However, these varieties don't provide the essential contents as naturally produced crop. These unnatural techniques spoil the soil. It all ends up in further environmental harm. Most of these unnatural techniques are wont to avoid losses.

But when the producers of the crops know the accurate information on the crop yield it minimizes the loss. Machine learning, a fast-growing approach that's spreading out and helping every sector in making viable decisions to create the foremost of its applications. Most devices nowadays are facilitated by models being analysed before deployment. The main concept is to increase the throughput of the agriculture sector with the Machine Learning models. Another factor that also affects the prediction is the amount of knowledge that's being given within the training period, as the number of parameters was higher comparatively. The core emphasis would be on precision agriculture, where quality is ensured over undesirable environmental factors.

So as to perform accurate prediction and stand on the inconsistent trends in the season a crop is chosen for cultivation and the area under cultivation various Machine Learning models like Random Forest Regression Model, Linear Regression, Ridge, Lasso, Elastic Net, Decision Tree Regression are applied to urge a pattern. By applying these machine learning models, I came into a conclusion that Random Forest algorithm provides the foremost accurate value. System predicts crop prediction from the gathering of past data. Using the past information on Year, Area, State and UT the major two crops Rice and Wheat was cultivated the model was trained to give meaningful prediction of production in tonnes.

Chapter 2

Preliminaries

2.1 Basic Definitions and Results

Definition 2.1.1: Machine Learning (ML)

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Definition 2.1.2: Algorithm

An algorithm is a computer program that adapt and evolve based on the data it is processed to produce predetermined outcomes.

Definition 2.1.3: Model

A model is a program that can find patterns or make decisions from a previously unseen dataset.

Definition 2.1.4: Predictor Variable (x)

A predictor variable is a variable or set of features whose values will be used to predict the value of the target variable.

Definition 2.1.5: Response Variable (Y)

It is the feature or the output variable that needs to be predicted by using the predictor variable. It is that feature of the dataset about which you want to gain a deeper understanding.

Definition 2.1.6: Training Data

Training data is the data you use to train an algorithm or machine learning model to predict the outcome you design your model to predict.

Definition 2.1.7: Testing Data

An unseen data used to test your model is called testing data, and you can use it to evaluate the performance and progress of your algorithms training and adjust or optimize it for improved results.

2.2 Machine Learning Process

The machine learning process involves constructing a predictive model to address a specific problem. The process typically consists of seven key steps:

1) **The Objective of the Problem**
   Clearly defining and understanding the problem statement that needs to be solved.

2) **Data Gathering**
   Identifying and selecting the appropriate sources from which to obtain the necessary data. This step aims to collect the data required for training the model.

3) **Preparing Data**
   Manipulating and organizing the data in a suitable format for model training. This step involves data cleaning tasks such as removing duplicates, correcting errors, handling missing values, normalizing data, and converting data types.

4) **Data Exploration**
   Visualizing the data to identify relevant relationships between variables, detect class imbalances, or perform other exploratory analyses.

Additionally, splitting the data into training and evaluation sets is done in this step.

5) Building a Model

Developing a predictive model that performs better than a baseline or reference model. This step involves selecting an appropriate algorithm or model architecture and training it using the prepared data.

6) Model Evaluation and Optimization

Assessing the performance of the trained model and optimizing its effectiveness. This includes scaling up the model to detect potential overfitting issues, regularizing the model to improve generalization, and tuning model parameters for better performance.

7) Predictions

Applying the trained model to new, unseen data (test set) that was previously withheld from the model. This data, for which the class labels are known, is used to evaluate the model's performance and obtain a better estimate of how the model will perform in real-world scenarios.

2.3 Ensemble Learning

Ensemble learning is a machine learning technique that combines multiple models to improve their performance. It reduces variance and bias of the model and improves its accuracy and robustness. It is an approach in which two or more models are fitted to the same dataset, and the predictions of each model are combined.

There are mainly three types of Ensemble learning methods -

1) Bagging -

It stands for Bootstrap Aggregating and involves training multiple models on different subsets of the data and combining their predictions. The goal is to reduce the variance of the model and prevent overfitting.

2) Boosting –

It involves training multiple models sequentially and giving more weight to the misclassified data. The goal is to improve the accuracy of the model and reduce the bias. Involves ML model like XGBoost.

3) Stacking -

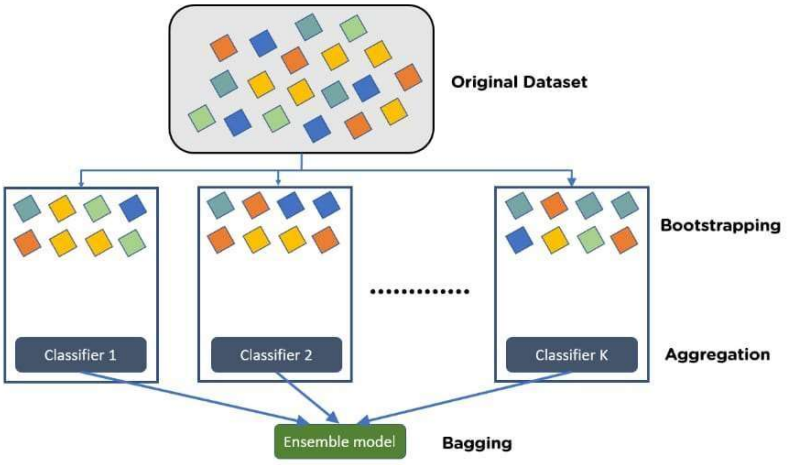
It involves training multiple models and using their predictions as input to a meta-model. The goal is to improve the accuracy and reduce the variance of the model by combining the strengths of multiple models. Involves methods to build new prediction models such as KNN (K Nearest Neighbour) or SVM (Support Vector Machine).

2.3.1 Bagging

Bagging is a machine learning technique that involves training multiple models on different subsets of the data and combining their predictions. The goal of bagging is to reduce the variance of the model and prevent overfitting. This method is commonly used to reduce variance and maintain bias.

The name "bagging" comes from the fact that the technique involves creating random subsets of the training data, also known as "bags". Each bag contains a random sample of the training data, and a model is trained on each bag.

The models can be of different types, but decision trees are often used because they tend to overfit the training data. By training multiple trees on different subsets of the data, bagging can help to reduce the variance of the model and improve its accuracy. Once the models are trained, their predictions are combined to make a final prediction. This can be done by taking the average of the predictions for regression tasks, or by using a voting scheme for classification tasks.



Machine learning algorithms that use the bagging method are as follows –

1) Random Forest -
   It is an ensemble learning method that combines multiple decision trees to make predictions. It uses bagging to reduce the variance of the model and improve its accuracy.

2) Bagged Decision Trees -
   It involves training multiple decision trees on different subsets of the data and combining their predictions. It is a type of bagging method that can be used for both regression and classification tasks.

3) Bagged SVM –
   It involves training multiple support vector machines (SVMs) on different subsets of the data and combining their predictions. It is a type of bagging method that can be used for classification tasks.

4) Bagged Neural Networks –
   It involves training multiple neural networks on different subsets of the data and combining their predictions. It is a type of bagging method that can be used for both regression and classification tasks.

All of these algorithms use the bagging method to reduce the variance of the model and improve its accuracy.

2.4 Fitted Models -

2.4.1 Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

The diagram above shows the structure of a Random Forest. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. To get a better understanding of the Random Forest algorithm, let's walk through the steps:

1) Pick at random k data points from the training set.
2) Build a decision tree associated to these k data points.
3) Choose the number N of trees you want to build and repeat steps 1 and 2.
4) For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all the predicted y values.

A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with nonlinear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.

Advantages -

1) Random forest regression is a powerful and flexible modelling technique that can handle complex data structures and high dimensional data.
2) It is less prone to overfitting than other modelling techniques, such as decision trees, due to the use of bagging and random feature selection.
3) It can handle missing data and categorical variables without the need for data pre-processing.
4) It provides feature importance measures that can be used to identify the most important predictors in the model.

Disadvantages -

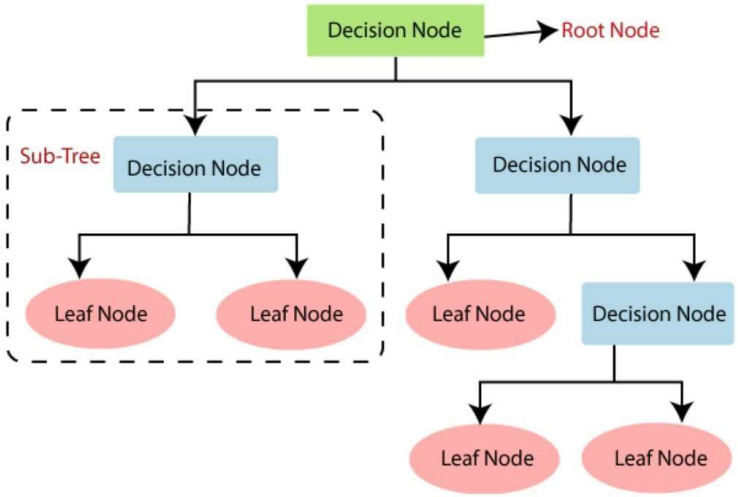
1) Random forest regression can be computationally expensive and require a lot of memory, especially for large datasets.
2) It is not as interpretable as simpler models, such as linear regression.
3) It may not perform well on small datasets or datasets with a small number of predictors.
4) It may not work well if there are strong correlations between predictors.

2.4.2 Decision Tree Regression

A Decision Tree is a predictive model that uses a set of binary rules in order to calculate the dependent variable. Each tree consists of branches, nodes, and leaves.

Terminologies used in this regression are as follows:

1) The root node represents the entire population and is divided into two or more homogeneous sets.
2) A decision node is when a sub-node splits into further sub-nodes.
3) A leaf is when a node does not split. These are also referred to as "Terminal Nodes".



A Decision Tree imposes a series of questions to the data, each question narrowing possible values, until the model is trained well to make predictions. These questions are determined completely by the model, including their content and order, and are asked in a True/False form. Decision trees are classified supervised learning model. It learns based on a known set of input data with known responses to the data. It is able to make a prediction by running through the entire tree, asking true/false questions, until it reaches a leaf node. The final prediction is given by the average of the value of the dependent variable in that leaf node.

Advantages -

1) Decision Tree Regression is a non-parametric algorithm that can model complex relationships between input and output variables without making assumptions about the underlying distribution of the data.

2) It is easy to interpret and visualize, making it a useful tool for understanding the patterns and relationships in the data.
3) It can handle both categorical and continuous input variables and can perform variable selection automatically.
4) It can handle missing values and outliers in the data.
5) It can be used for both classification and regression problems.

Limitations -

1) Decision Tree Regression is prone to overfitting, especially when the tree is deep and the training data is small.
2) It may not perform well when the input variables have complex interdependencies or when there are many irrelevant variables in the data.
3) It may not generalize well to new data if the training data is not representative of the entire population.

2.4.3 Ridge Regression

Ridge regression is a technique that estimates the coefficients of multiple regression models when the independent variables are highly correlated or multicollinear. It is a form of regularized regression that adds regularization to the least square problem. It can be used to predict continuous values and to reduced overfitting.

For any type of regression machine learning model, the usual regression equation forms the base which is written as: $Y = X\beta + e$

Where,

Y is the dependent variable,

X represents the independent variables,

β is the regression coefficients to be estimated, and

e represents the errors are residuals.

Standardization -

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must

somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.

Assumptions -

The assumptions of ridge regression are the same as that of linear regression: linearity, constant variance, and independence.
To avoid overfitting the learning should constrain the solution in order to fit a global pattern. Adding such a penalty will force the coefficients to be small, i.e., to shrink them toward zeros.

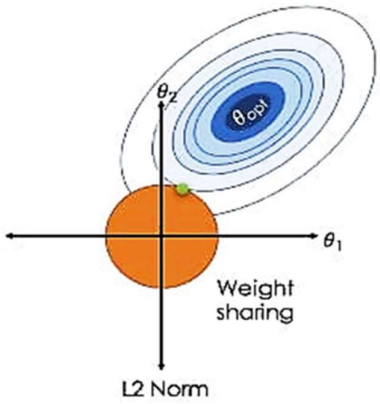
Ridge regression impose a l penalty on the coefficients, i.e., it penalizes with the Euclidean norm of the coefficients while minimizing SSE. The objective function becomes:

$$Ridge(\theta) = \|y - x\theta\| + \lambda\|\theta\|$$

Mathematically it can be stated as:

Minimize: $\|y - X\theta\|_2^2$

Subject to: $\|\theta\|_2^2 \leq k$



Advantages -

1) Ridge regression can help to prevent overfitting in linear regression models by adding a regularization term to the loss function.
2) It can handle multicollinearity in the data by shrinking the coefficients of correlated predictors towards each other.
3) It is computationally efficient and easy to implement.
4) It can improve the stability and generalizability of the model.

Disadvantages -

1) Ridge regression assumes that all predictors are equally important, which may not be the case in some situations.
2) It may not perform as well as other regularization techniques, such as Lasso regression, when there are a large number of predictors.
3) It may not work well if the data is not well-suited to linear modelling.
4) It may require some tuning of the regularization parameter to get optimal results.

2.4.4 Lasso Regression

The word "LASSO" stands for Least Absolute Shrinkage and Selection Operator. It is a statistical formula for the regularisation of data models and feature selection.

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso regression penalizes the coefficients by the l1 norm. This constraint will the capacity of the learning algorithm.

To add such a penalty forces the coefficients to be small, i.e., it shrinks them toward zero. The objective function to minimize becomes:

$$lasso(\theta) = \|y - x\theta\| + \lambda\|\theta\|$$

If a regression model uses the L1 Regularization technique, then it is called Lasso Regression.

Mathematical equation of Lasso Regression:

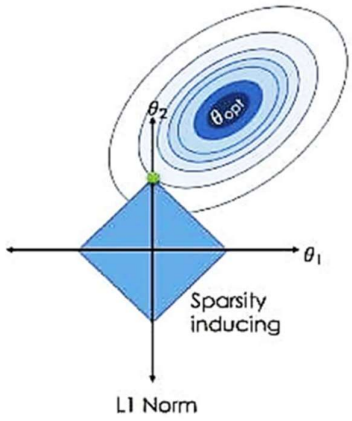
Residual Sum of Squares + λ × (Sum of the absolute value of the magnitude of coefficients) where, λ denotes the amount of shrinkage.

• λ = 0 implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model
• λ = ∞ implies no feature is considered i.e., as λ closes to infinity it eliminates more and more features
• The bias increases with increase in λ and Variance increases with decrease in λ

Mathematically it can be stated as:

Minimize: $\|y - X\theta\|_2^2$

Subject to: $\|\theta\|^1 \leq k$



Advantages -

1) Lasso regression can help to prevent overfitting in linear regression models by adding a regularization term to the loss function.
2) It can perform feature selection by setting the coefficients of irrelevant predictors to zero.
3) It can handle multicollinearity in the data by shrinking the coefficients of correlated predictors towards each other. 4) It is computationally efficient and easy to implement.

Disadvantages -

1) Lasso regression may not perform as well as other regularization techniques, such as Ridge regression, when there are a large number of predictors.
2) It may not work well if the data is not well-suited to linear modelling.
3) It may require some tuning of the regularization parameter to get optimal results.
4) It may not be able to handle situations where there are more predictors than observations.

2.4.5 Elastic Net Regression

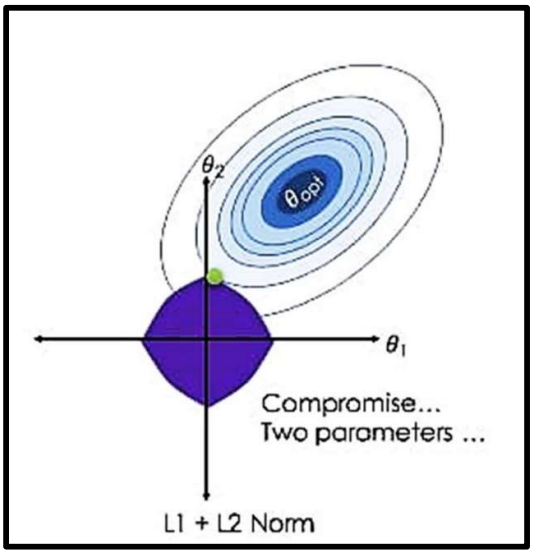
Elastic net (also called ELNET) regression is a statistical hybrid method that combines two of the most often used regularized linear regression techniques, lasso, and ridge, to deal with multicollinearity issues when they arise between predictor variables. Regularization aids in solving the overfitting issues with the models.

It is also used for regularizing and choosing the essential predictor variables that significantly impact the response variable. Ridge employs an L2 penalty, while lasso employs an L1. Since the elastic net utilizes both the L2 and the L1 models, the question of choosing between either one does not arise.

An elastic net is a combination of two regressions, lasso, and ridge, and hence the resultant equation to calculate it is:

$$\varepsilon_{mse} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(\theta_0 + \theta_1 X_1 + \cdots\cdots \theta_n X_n)\right)^2 + \gamma \sum_{i=0}^{n}|\theta_i| + \sigma \sum_{j=1}^{n}\theta_i^2$$

The elastic net penalty comes in two varieties: l1 and l2. The lasso and ridge regression models are two types of regularization models that apply l1 and l2 penalties, respectively. The absolute value of the coefficient's magnitude is added as a penalty term in the lasso regression model. The ridge regression adds the coefficient's squared magnitude as a penalty to the loss function.



Thus, it deals with both multicollinearity problems and the selection of regression coefficients. ELNET uses coefficient regression shrinkage towards zero or equal to zero to reduce the occurrence of predictor variables. The tuning parameter (λ1) multiplied by the sum of coefficient variables (l1 norms) absolute values is utilized for this purpose. The tuning parameter (λ2) is multiplied by the l2 norm's squared coefficient variables to handle the high correlation between the predictor variables. The ELNET regression approach helps produce a fitting, interpretable model by minimizing unnecessary variables that do not appear in the final model to improve prediction accuracy. The ELNET manages multicollinearity by maintaining or excluding highly correlated predictor variables from the fitted model. While building ElasticNet regression model, both hyperparameters (L2) and (L1) need to be set.

ElasticNet takes the following two parameters:

- alpha – Constant that multiplies the penalty terms. Default value is set to 1.0.
(alpha = $\gamma + \sigma$)
- l1_ratio – The ElasticNet mixing parameter, with $0 \leq$ l1_ratio ≤ 1.

$$l1\_ratio = \frac{\gamma}{\gamma+\sigma}$$

Where, l1_ratio = 0 implies that the penalty is an L2 penalty.

l1_ratio = 1 implies that it is an L1 penalty.

$0 <$ l1_ratio < 1 implies that the penalty is a combination of L1 and L2.

Chapter 3

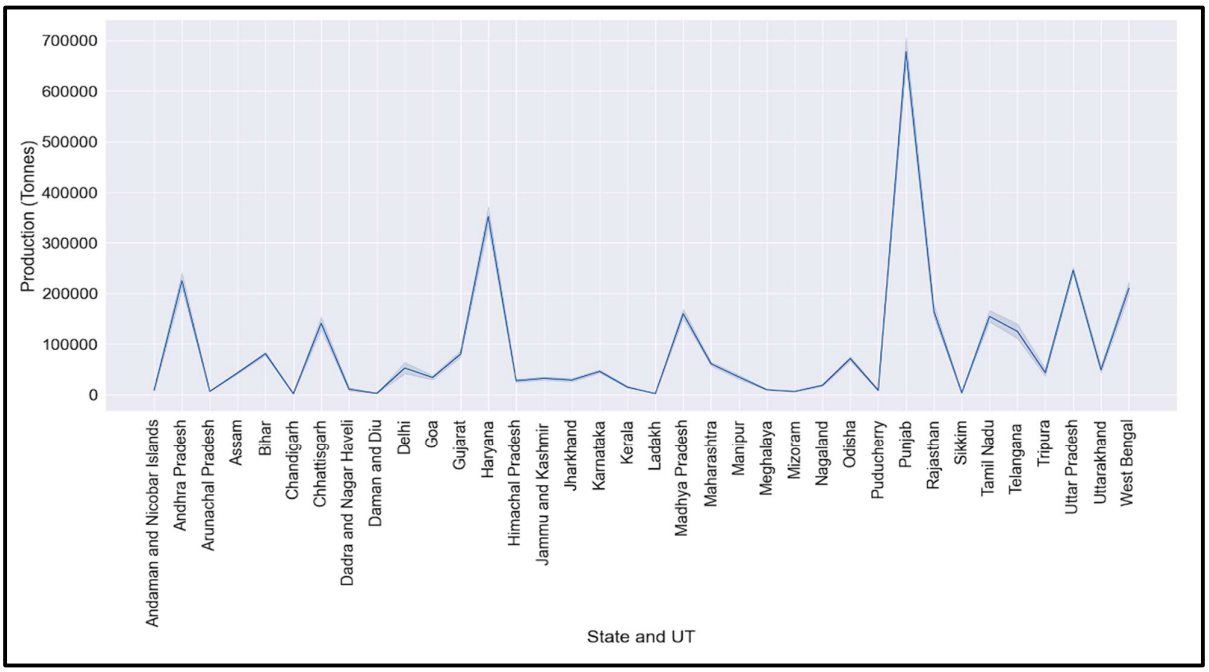
Analysis of crop production in India

India is a global agricultural powerhouse and is unquestionably the largest livelihood provider in India, more so vast in rural areas. Being that important sector, it requires good quality of analytical view and better farming practices to avoid damage and loss on the part of our hardworking farmers. For our examination approach crop production details state-wise which then goes a level deep with district details was extracted from a government owned website, https://www.aps.dac.gov.in/. The data from 1997-2020 with major crops Rice and Wheat is taken under study. Some other features that are available to us are the crop season and area under cultivation.

This data is only about India. And so, the scope of research is also going to be limited to just India.

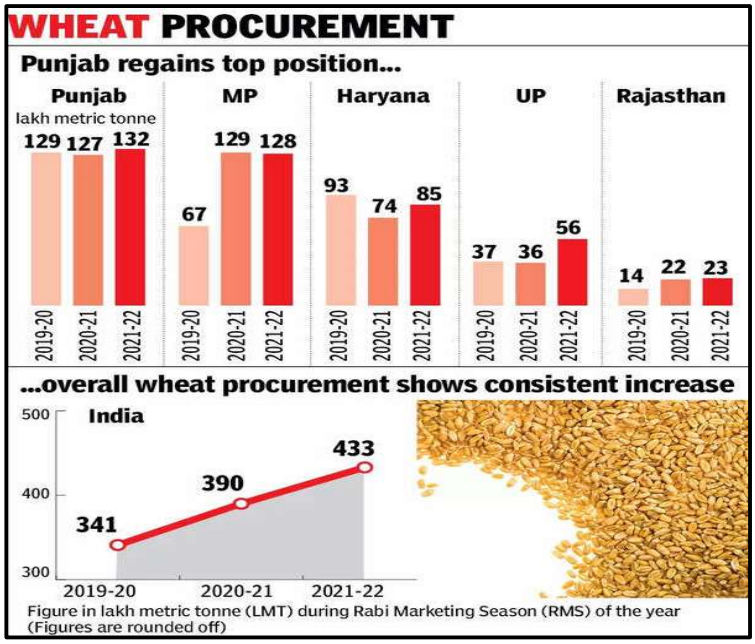
3.1 State and District-wise Production analysis (1997 – 2020)

3.1.1 Graph representing State and Union Territories Vs Production in tonnes

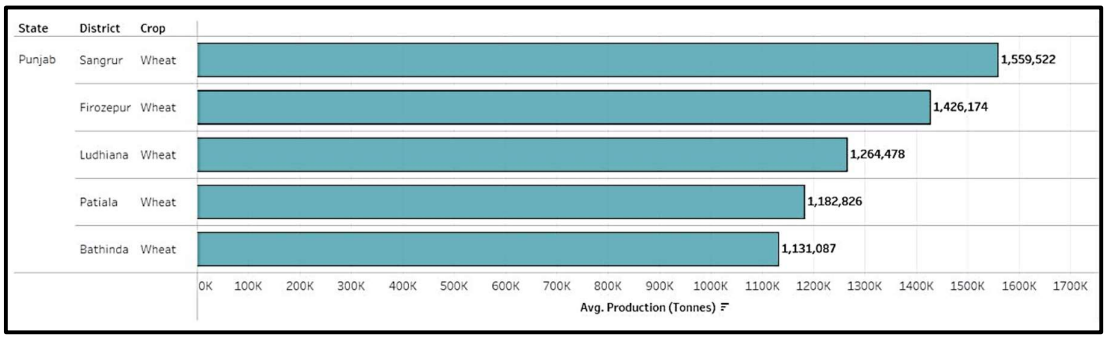


3.1.2 Wheat production –

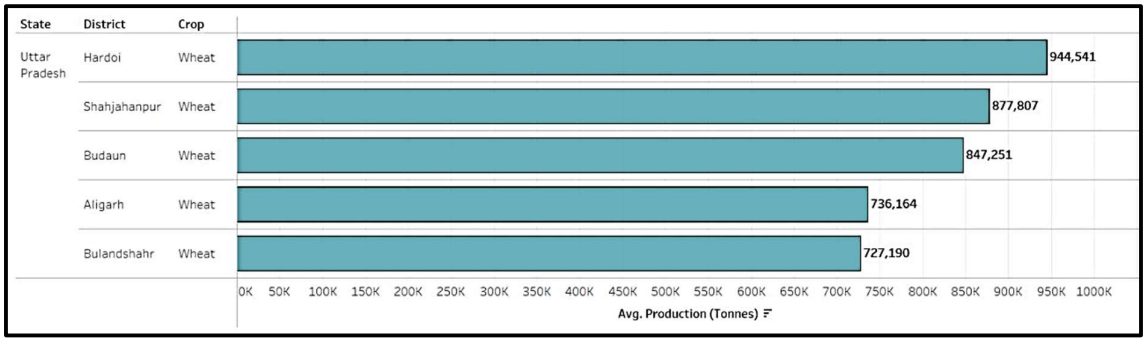
According to the procurement report of 2022 by Times Of India (TOI),



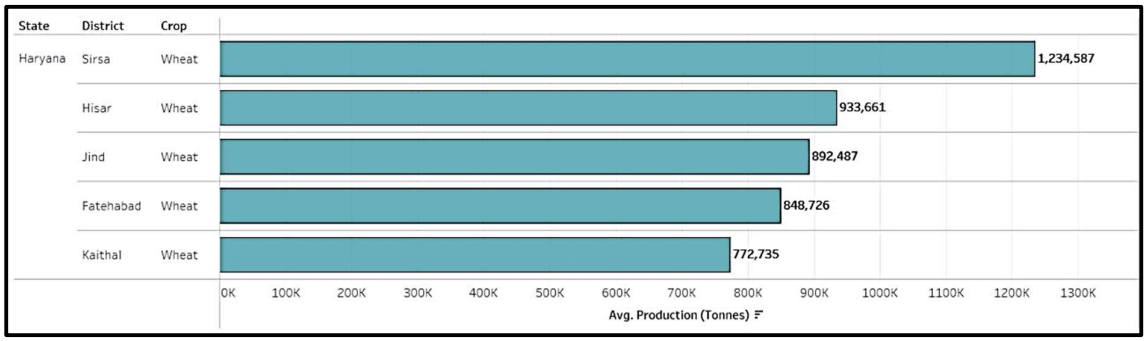
i. Punjab: As per the current trends, Punjab is the largest producer of wheat in India, accounting for around 40% of the total wheat production in the country. On analysis, the state shows up as the largest producer with major wheat-growing districts in the state including Sangrur, Firozepur, Ludhiana, Patiala and Bhatinda. The climate of Punjab is well suited and the soil in is also rich in nutrients, which provides the necessary conditions for wheat. In addition, the state has a well-developed irrigation system.



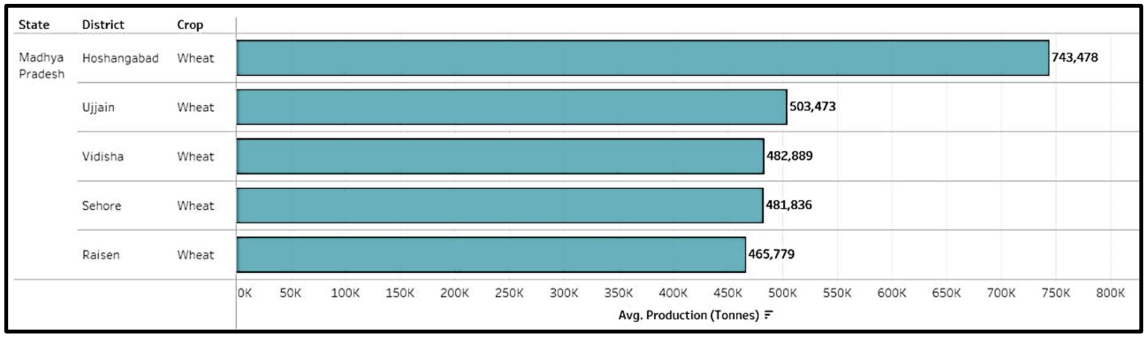
ii. Uttar Pradesh: Uttar Pradesh is the second-largest producer of wheat in India currently, accounting for around 18% of the total wheat production in the country. The state is known for its high-yielding wheat varieties and favourable agro-climatic conditions. The past 22 years analysis shows that the state stands up as the second largest for wheat production in India. The major wheat-growing districts in Uttar Pradesh include Hardoi, Shahjahanpur, Budaun, Aligarh and Saharanpur.



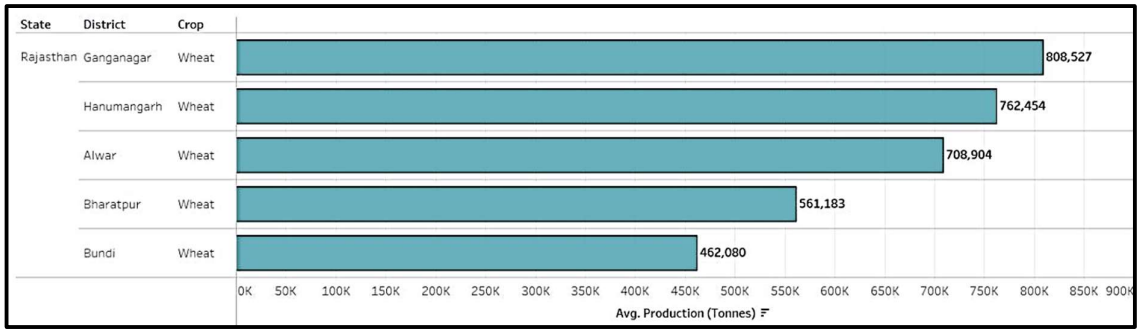
iii. Haryana: The third-largest producer of wheat in India, accounting for around 15% of the total wheat production in the country. Land of Haryana is flat, covered with loamy soil which is very suitable for agriculture. According to our analysis over the data, major wheat-growing districts in Haryana include Sirsa, Hisar, Jind, Fatehabad and Ambala.



iv. Madhya Pradesh: The state is said to be the fourth-largest producer of wheat in India right now, accounting for around 10% of the total wheat production in the country. The state has a favourable climate and fertile soil for wheat cultivation. So, the analysis over the data concludes the same and the major wheat-growing districts in the state are Hoshangabad, Ujjain, Vidisha, Sehore and Raisen.

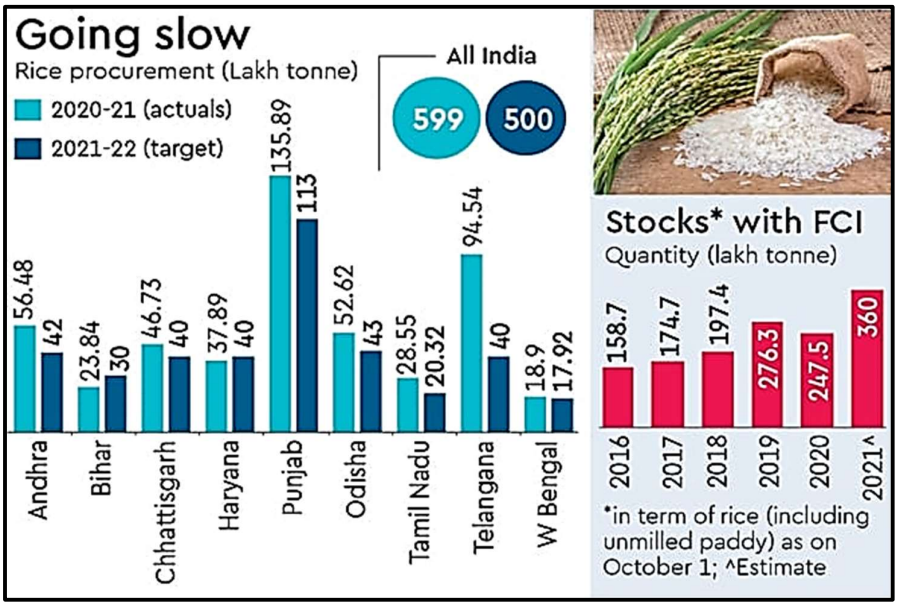


v. Rajasthan: The state is the fifth-largest producer of wheat in India currently, accounting for around 8% of the total wheat production in the country. The Major wheat-growing districts in Rajasthan include Ganganagar, Hanumangarh, Alwar, Bharatpur and Kota. The state is better is terms of wheat production because its desert area has a potential.

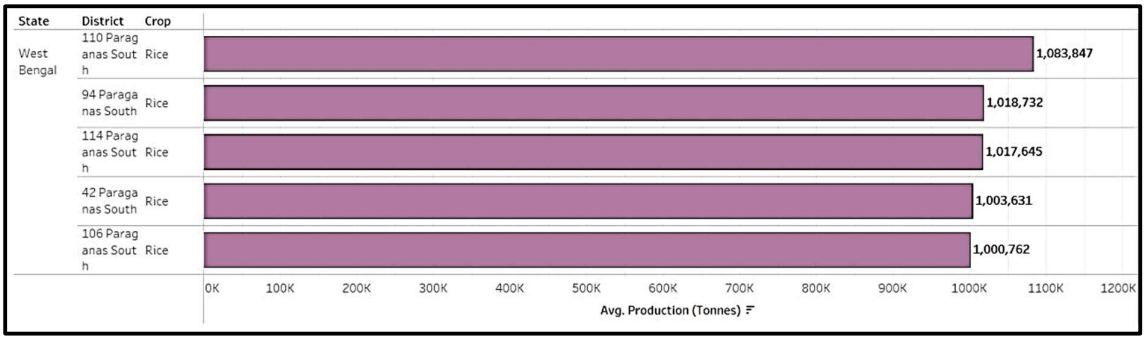


3.1.3 Rice production –

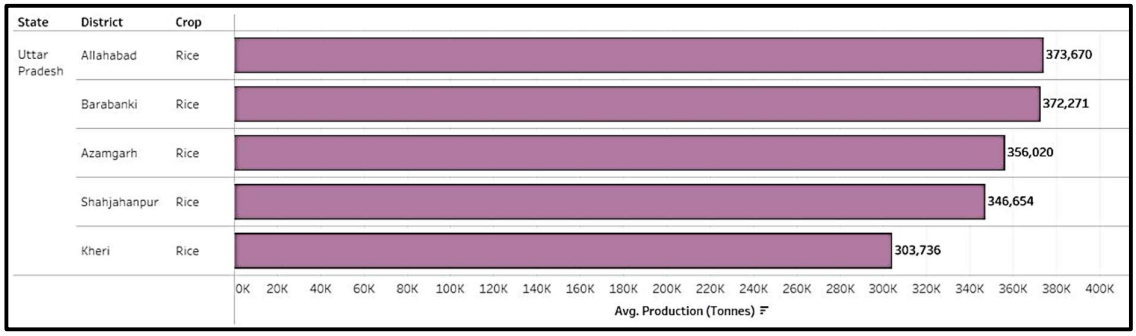
According to the procurement analysis report by Food Corporation of India (FCI),



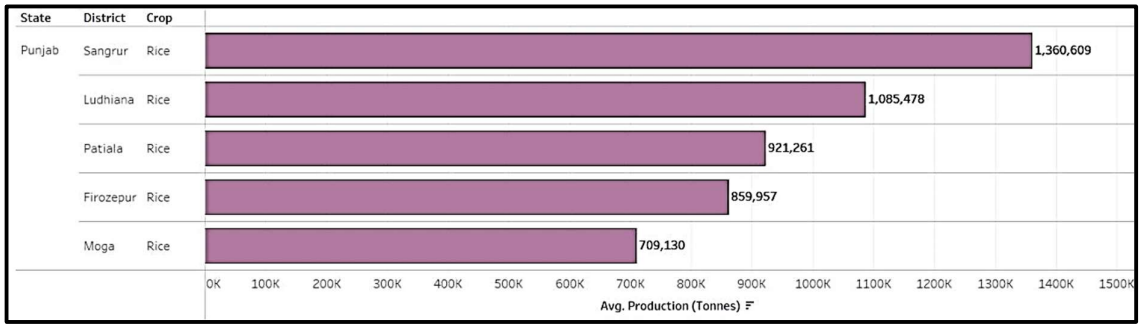
i. <u>West Bengal</u>: The state is the largest producer of rice in India, accounting for around 13.95% of the country's rice. The state has a favorable climate and fertile soil for rice cultivation. Major rice-growing districts in West Bengal include Burdwan, Hooghly, and Nadia. Rice cultivation requires an ample of water and a type of soil which holds the water, West Bengal have such type of soil.



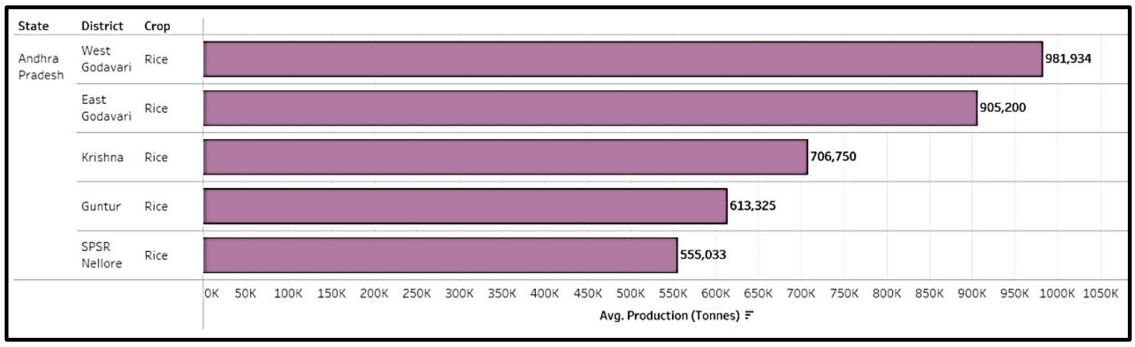
ii. Uttar Pradesh: The second-largest producer of rice in India, accounting for around 13% of the total rice production in the country. The state has a favorable climate and fertile soil for rice cultivation. About 25% of the cultivated area of the state is devoted to rice production. Also, Allahabad, Barabanki, Gorakhpur etc., are the major rice-producing districts of the state. In Uttar Pradesh, you will find Jaya, Panth-4, Mahsuri, Pusa Basmati Rice and Kasturi Basmati Rice. Farmers also cultivate the rice in an area of 59 hectares in UP.



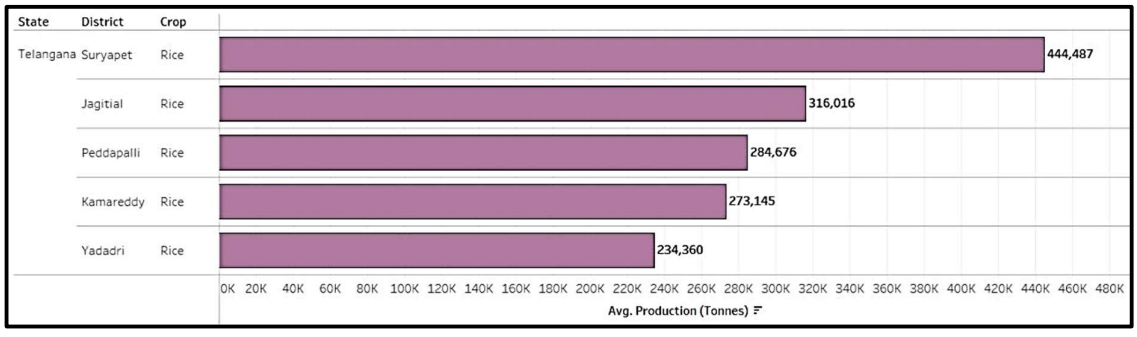
iii. Punjab: The state is the third-largest producer of rice in India, accounting for around 12% of the total rice production in the country. The state has a favorable climate and fertile soil for rice cultivation. Major rice-growing districts in Punjab include Sangrur, Ludhiana, Patiala, Firozpur and Moga. This state is a predominantly wheat-growing state, farmers have also used crop rotation and perennial irrigation sources to grow rice.



iv. Andhra Pradesh: It is the fourth-largest producer of rice in India, accounting for around 11% of the total rice production in the country. The state has a favorable climate and fertile soil for rice cultivation. Major rice-growing districts in Andhra Pradesh include East Godavari, West Godavari and Krishna. Also, the state is renowned as the 'Rice Bowl of India' as Krishna-Godavari delta region is historically as know by the same name.



v. Telangana: Telangana is the fifth-largest producer of rice in India, accounting for around 9% of the total rice production in the country. The state has a favorable climate and fertile soil for rice cultivation. Major rice-growing districts in Telangana include Suryapet, Jagitial, and Peddapalli.

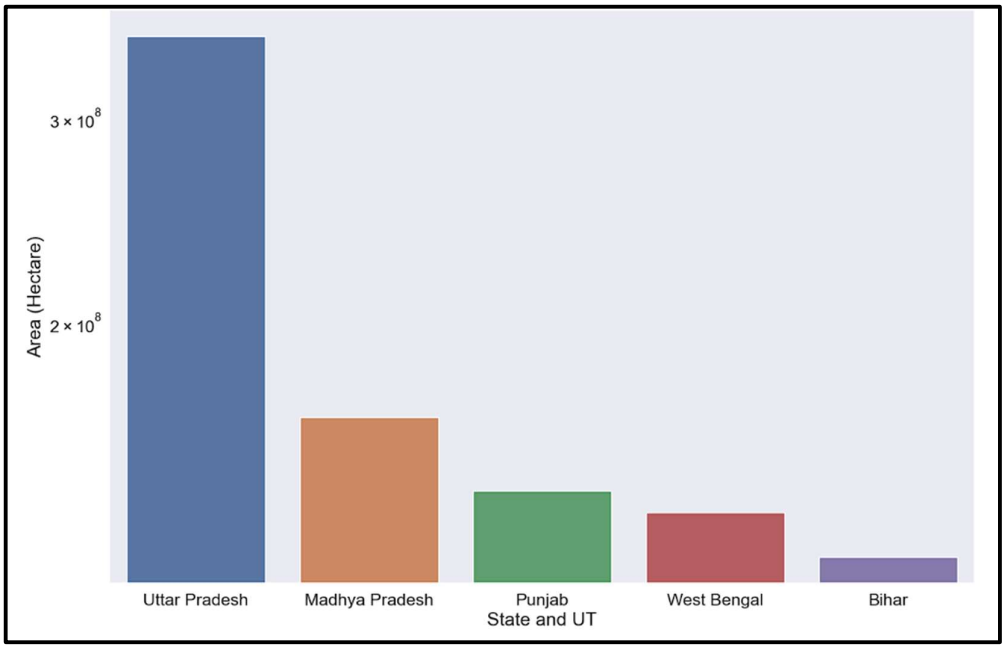


3.2 Agricultural area under cultivation Vs Production in tonnes

Agricultural area and production are closely related in India, where agriculture is a major source of livelihood for millions of people.

The amount of agricultural area available for cultivation is a key factor that determines the overall production levels of crops such as rice and wheat. In general, states with larger areas under cultivation for these crops tend to have higher production levels, while states with smaller areas under cultivation may have to rely on higher yields per hectare to achieve similar production levels.

According to the Ministry of Agriculture and Farmers' Welfare, the state with the highest area under cultivation in India is Uttar Pradesh, with around 25 million hectares of area used for agriculture. As per the analysis on the past data, the state is the leading one using the land for agriculture purpose.



3.3 Crop Vs Production in tonnes

Over the past few decades, India has seen significant growth in both rice and wheat production. According to the United States Department of Agriculture, the total rice production in India has increased from around 80 million tonnes in 1997 to around 120 million tonnes in 2020-21. Similarly, the total wheat production in India has increased from around 60 million tonnes in 1997 to around 109 million tonnes in 2020-21. So, both rice and wheat production have increased significantly over the past few decades.

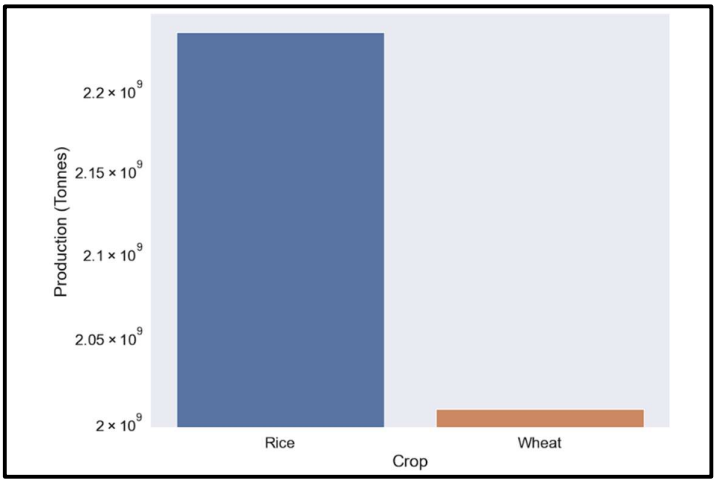
Also, the Rice is generally grown in greater quantities than wheat in India. According to the Ministry of Agriculture and Farmers' Welfare, in 2020-21, the total area under rice cultivation in India was around 43 million hectares, while the total area under wheat cultivation was around 30 million hectares. So, rice is grown more than wheat in India.

CODE -

```
DF = df.copy()
DF = DF.groupby(
    by='Crop')['Production (Tonnes)'].sum().reset_index().sort_values(
        by='Production (Tonnes)', ascending=False)

fig, ax = plt.subplots(figsize=(10, 8))
sns.barplot(x=DF['Crop'], y=DF['Production (Tonnes)'], errwidth=2)
sns.set(font_scale=1.5)
plt.yscale('log')
DF
```

GRAPHICAL REPRESENTATION,



3.3.1 Rice

3.3.1.1 State, District-wise production of rice

As per the data from the Ministry of Agriculture and Farmers' Welfare, the top rice-producing states in India from 1997 to 2020 were West Bengal, Uttar Pradesh, and Andhra Pradesh. West Bengal consistently maintained its position as the largest producer of rice in India during this period, with a production of around 14-15 million metric tons per year. Uttar Pradesh and Andhra Pradesh were also among the top producers during this period, with production levels ranging from 10-14 million metric tons per year.

Top 5 districts with highest yield of crop in West Bengal,

CODE -

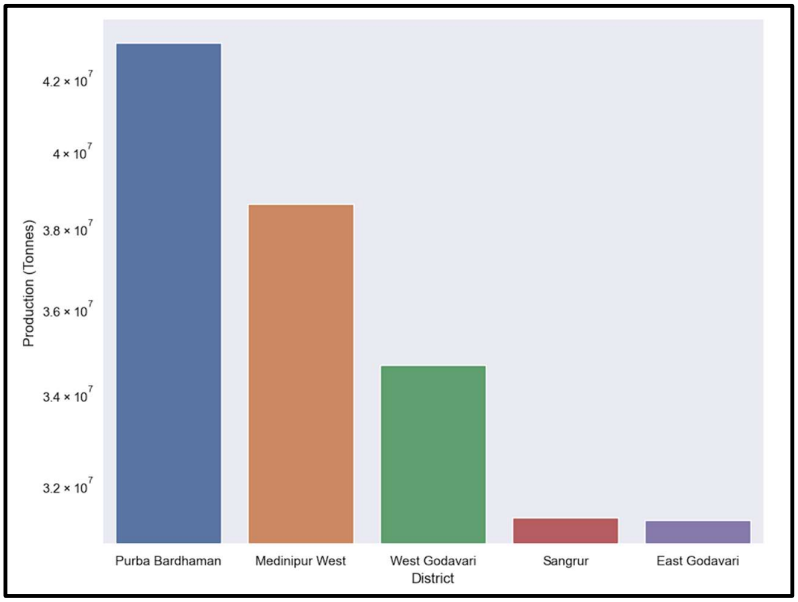
```
DF = rice.copy()
DF = DF.groupby(
    by='District')['Production (Tonnes)'].sum().reset_index().sort_values(
        by='Production (Tonnes)', ascending=False)


fig, ax = plt.subplots(figsize=(10, 8))
sns.barplot(x=DF['District'].head(), y=DF['Production (Tonnes)'].head(), errwidth=2)
sns.set(font_scale=1)
plt.yscale('log')
DF.head()
```

OUTPUT -

	District	Production (Tonnes)
622	Purba Bardhaman	42976070.0
538	Medinipur West	38585256.0
796	West Godavari	34650471.0
671	Sangrur	31294000.0
312	East Godavari	31229411.0

GRAPHICAL REPRESENTATION,

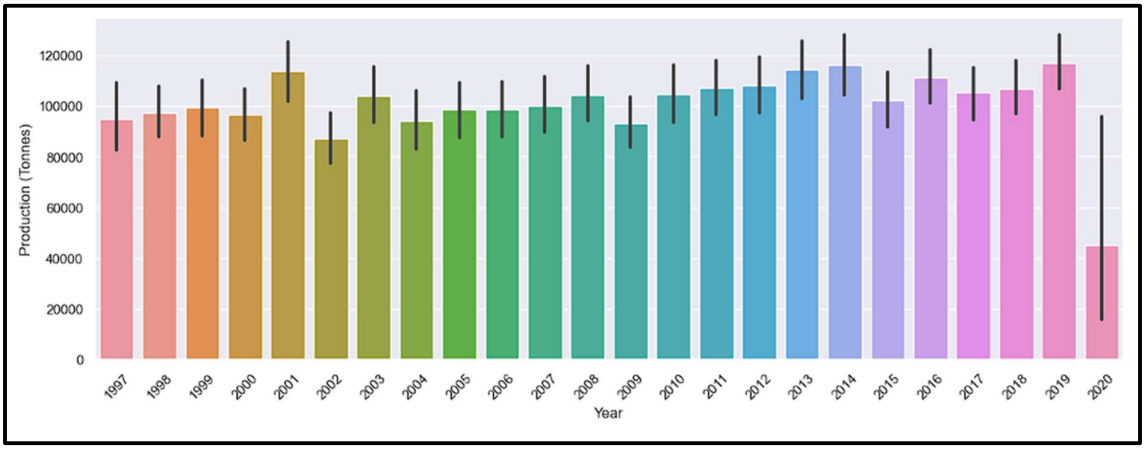


3.3.1.2 Year-wise production of rice

CODE -

```
plt.figure(figsize=(15,5))
sns.barplot(x = 'Year', y = 'Production (Tonnes)', data = rice)
plt.xticks(rotation=45)
plt.show()
```

GRAPHICAL REPRESENTATION,



3.3.2 Wheat

3.3.2.1 State, District-wise production of wheat

The top wheat-producing states in India from 1997 to 2020 were Uttar Pradesh, Punjab, and Haryana. Punjab is said to be the leading state.

Top 5 districts with highest yield of crop in Punjab,

CODE -

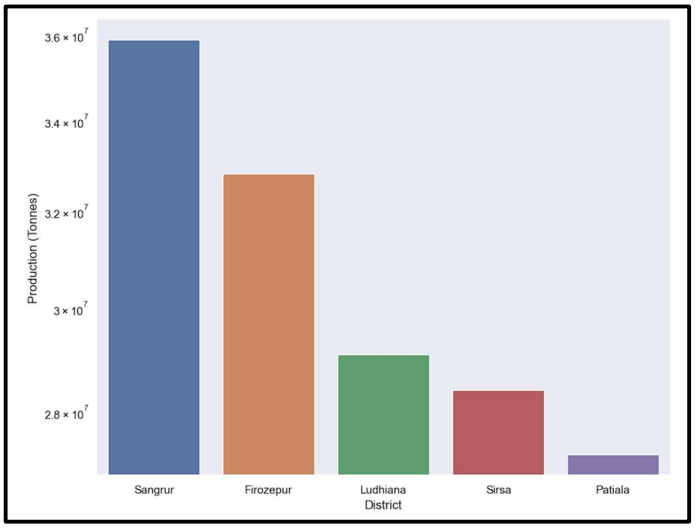
```
DF1 = wheat.copy()
DF1 = DF1.groupby(
    by='District')['Production (Tonnes)'].sum().reset_index().sort_values(
        by='Production (Tonnes)', ascending=False)


fig, ax = plt.subplots(figsize=(10, 8))
sns.barplot(x=DF1['District'].head(), y=DF['Production (Tonnes)'].head(), errwidth=2)
sns.set(font_scale=1)
plt.yscale('log')
DF1.head()
```

OUTPUT -

	District	Production (Tonnes)
537	Sangrur	35869000.0
230	Firozepur	32802000.0
389	Ludhiana	29083000.0
571	Sirsa	28395500.0
478	Patiala	27205000.0

GRAPHICAL REPRESENTATION,

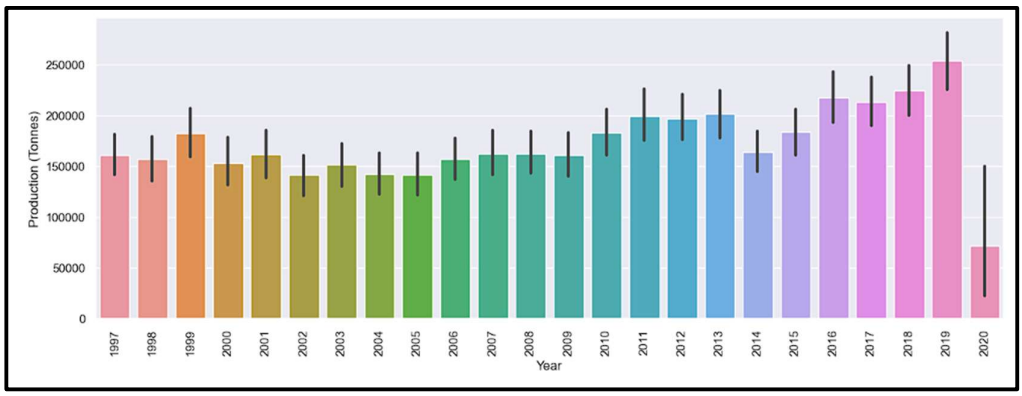


3.3.2.2 Year-wise production of wheat

CODE -

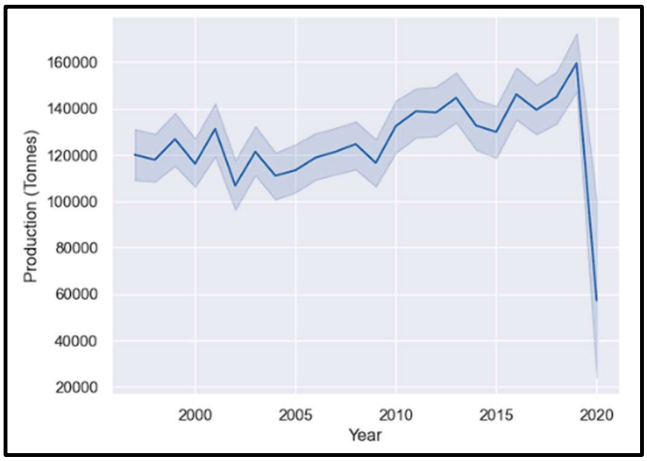
```
plt.figure(figsize=(15,5))
sns.barplot(x = 'Year', y = 'Production (Tonnes)', data = wheat)
plt.xticks(rotation = 90)
plt.show()
```

GRAPHICAL REPRESENTATION,



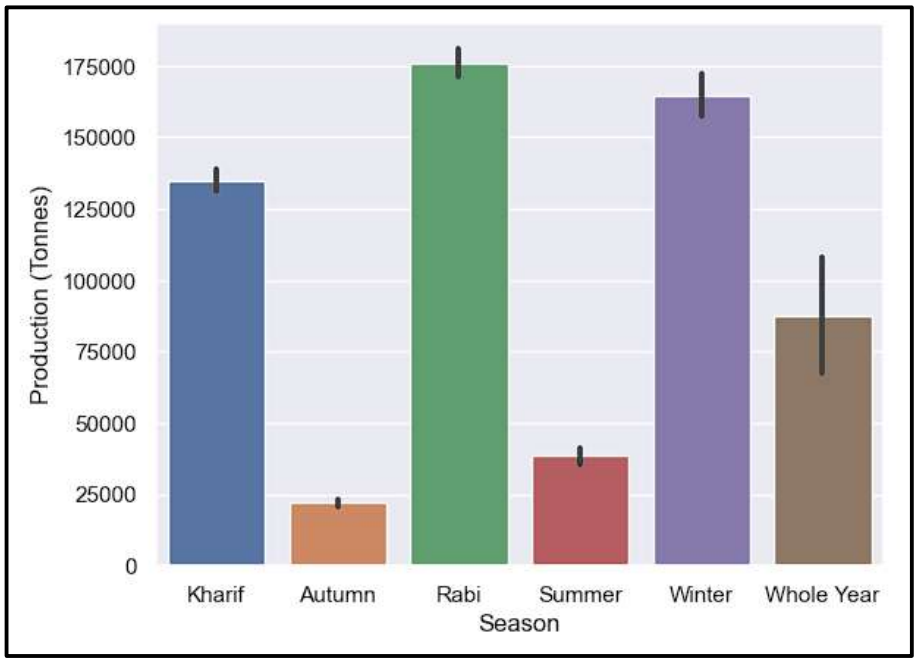
3.4 Year-wise Production in tonnes

After a fluctuating growth the production dropped down to around 2019. According to the reports, COVID-19 disrupted certain activities in agriculture and the supply chain as well. Few reports indicates that harvesting activities were suspended due to lack of migrant workers, particularly in Northwest India where wheat is cultivated.



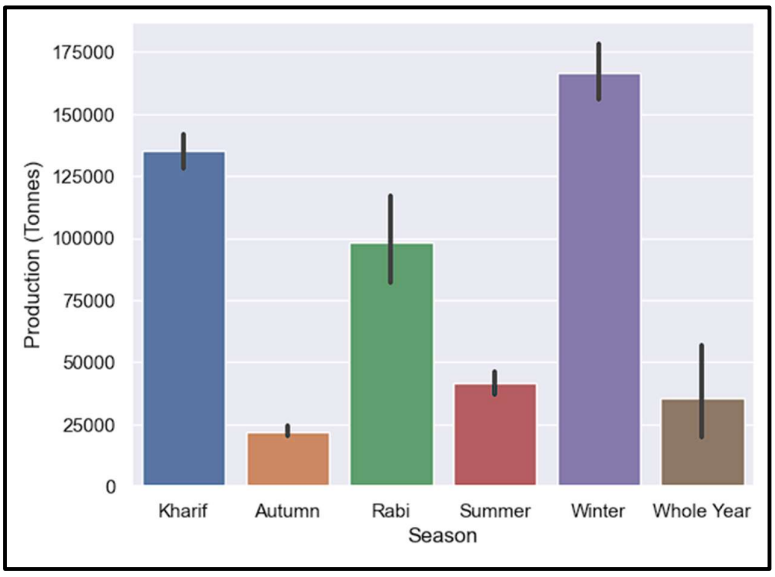
3.5 Season Vs Production

Rice and wheat are two of the most important food crops grown in India. Rice is mainly grown during the kharif season, while wheat is mainly grown during the rabi season. The production of these crops varies depending on the season, with different factors affecting the production during each season. Despite rest of the challenges like climate and pests, India has managed to maintain a steady increase in the production of rice and wheat, which has helped to meet the growing demand for food in the country.



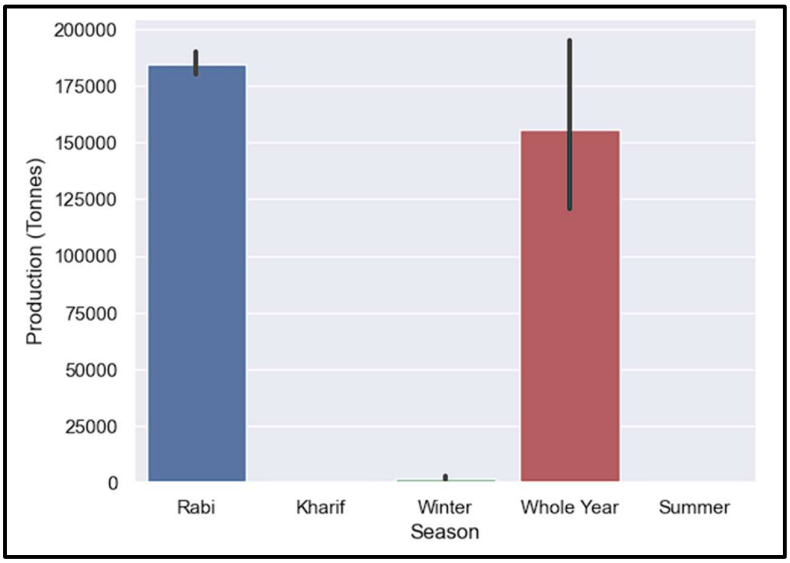
3.5.1 Season based production of rice

Being a kharif crop, rice is grown in another season as well. This might be possible due to the favourable climatic and soil conditions in certain regions.



3.5.2 Season based production of wheat

Wheat is a rabi crop but in some regions the production is certainly possible for the entire year as well.



Chapter 4

Prediction of Crop Production

I trained the data using Random Forest Regression model and predicted the crop production. I choose Random Forest regression algorithm as this was the best fit for the dataset under study, better predictions were made by this model as compared to the rest models. In this case n_estimators refer to the number of trees in the forest and I took it as 100 to avoid overfitting as Random Forest algorithm does not allow overfitting if there are more trees. Since the data under study was large enough and with many features, this algorithm has the ability to work upon it and deal with higher dimensionality as well.

```
1  from sklearn.ensemble import RandomForestRegressor
```

executed in 19ms, finished 00:06:11 2023-05-29

```
1  rfr = RandomForestRegressor(n_estimators = 100, random_state = 0)
2  rfr.fit(X_train, Y_train)
3  print('Model trained!')
```

executed in 1m 40.3s, finished 00:18:54 2023-05-29

```
Model trained!
```

This model is said to have 97% accurate in forecasting crop productions.

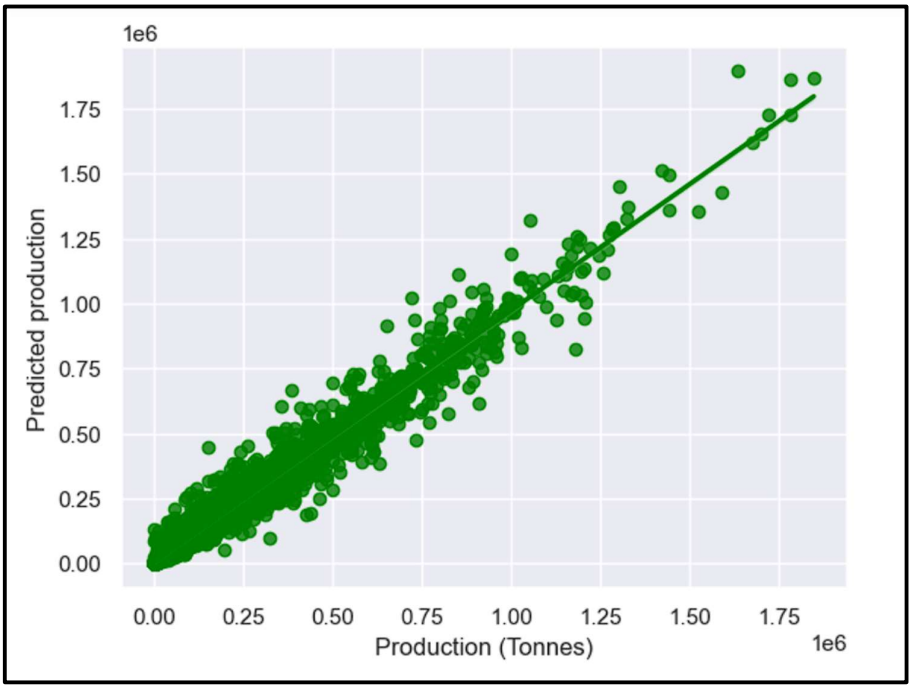
CODE -

```
1  Predicted = pd.DataFrame()
2  Predicted['Production_true'] = Y_test
3  Predicted['Production_predicted'] = rfr.predict(X_test)
4  Predicted.head(10)
```

OUTPUT -

	Production_true	Production_predicted
4404	10159.00	9413.68
3926	51151.00	63054.55
32516	452758.00	386498.07
11426	222000.00	246055.35
17286	9200.58	9737.27
9235	458453.00	448247.83
24097	250.00	233.40
9606	3461.00	3122.39
28157	3300.00	3726.03
35153	239884.00	238001.41

GRAPHICAL REPRESENTATION,



Chapter 5
Conclusion

Crop yield prediction is still remaining as a challenging issue for farmers. The aim of this research is to propose and implement a rule-based system to predict the crop yield production from the collection of past data. This has been achieved by applying association rule mining on agriculture data from 1997 to 2020.

Farmers require modern technologies to help them raise their crops. Agriculturists can be notified about accurate crop predictions on a timely basis. The agriculture factors were analysed using a variety of machine learning approaches. The features are chosen are determined by the dataset availability and research goal. According to studies, models with more characteristics may not always deliver the highest yield prediction performance. Models with more and fewer features should be evaluated to discover the best performing model. The findings reveal that while no definitive conclusion can be taken about which model is the best, they do show that some machine learning models are utilized more frequently than others. With the help of Random Forest Regression model, I got an accuracy of about 97%.

Scope of the project

I have taken into consideration only two major factors i.e., the area and season of the districts. There are other factors such as the fertility and type of soil present in the area which would affect the crop yield. Also, one of the major factor climatic conditions.

On proper analysis, I can help find the soil type and fertilizers and add on the weather conditions which can be helpful in maximizing and predicting the crop yield. P.C.A and many Deep Learning techniques can be used on images of the field and crop to detect if they are infected by any disease or the presence of weed, which also affects the quality of the crop, and can be separated from the healthy crops at the earliest possible.

Bibliography

[1] https://www.academia.edu/

[2] https://apeda.gov.in/apedawebsite/SubHead_Products/Wheat.html

[3] https://www.aps.dac.gov.in/

[4] https://agricoop.nic.in/sites/pocketbook

[5] https://edis.ifas.ufl.edu/publications

[6] https://techvidvan.com

[7] www.ijraset.com/research-paper/agriculture-crop-yield-prediction