Environmental Research 202 (2021) 111624

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Prediction of the production of crops with respect to rainfall

Benny Antony

Vellore Institute of Technology, India

ARTICLE INFO

Keywords:
Regression
Machine learning
Rice
Wheat

ABSTRACT

Agriculture is one of the most important sectors in the Indian context. It is one of the highest employing sectors in the Indian scenario. Unlike other sectors agriculture is highly dependent on the quality and the quantity of both the external factors like rainfall, climate, pH of the soil, fertilizers and insecticides used, and internal factors like the quality of seeds. This paper predicts the production of crops as a function of rainfall for four Indian States. This knowledge can be implemented in generating a rough overview of how the production is based on rainfall and how much can a specific crop production for the amount of rainfall it receives. Two crops each belonging to four different states are chosen and the best regression model for the crop of the state is chosen. There is no research done solely on how rainfall affects crops of particular states. The proposed method of evaluation is better than other existing methods of evaluation as it evaluates all the regression techniques (Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, Random Forest, and XGBRegression) for two crops of four individual states. For balanced evaluation, two states of North India and two states of South India are selected. The regression techniques are evaluated based on their Mean Squared Error.

1. Introduction

India is an agrarian economy. Developing countries have noted that their populations show seasonal weight fluctuations due to food shortages seen before the first harvest, which occurs late in the wet season. Excessive rain during short periods can cause flash floods. Climatic factors have more effect on productivity than the genotype of the crop (Khakhi and Wang, 2019). As of 2018 agriculture and its allied sectors account for almost 17–18 percent of GDP. It employs more than 50 percent of the Indian workforce. It exports more than 38 billion dollars as of 2013. The Indian population is more than 1.35 billion, food is one of the major driving forces of any population. This is evident in the Indian context as the majority of the crops grown in India are food crops, and foods for such a huge population are sourced locally making it economical than importing food crops. Of the many factors that agriculture one of the most important factors is rainfall. Prediction of the crop output with respect to the rainfall and the respective algorithm is carried out in this paper. Regression methods are preferred and chosen over neural networks as regression methods give a lower mean squared error than artificial neural networks. Regression methods with hyper-parameter fine-tuning perform better than neural networks.

However, freely available satellite imagery with low or moderate resolutions showed some limits in specific agricultural applications, e.g., where crops are grown by rows. Indeed, in this framework, the satellite's output could be biased by intra-row covering, giving inaccurate information about crop status. Decision Tree Learning is used for splitting the sample space using a recursive algorithm for formulation into a simple model. Random forest is used to generate multiple trees of sub-sampled features and functions as a high dimensional input predictor. Gradient boosted decision tree is used to build a unionistic predictive model on the reweighted data with the conjunction of the decision tree (Boukhris et al., 2020). The contribution presents a concept for a comprehensive artificial intelligence (AI) system which spans the whole farming value chain and operates on independent AI modules that are interconnected by a comprehensive cloud network. Data mining experiments were done for the proving of the proof of concept of the better algorithmic value of Multivariate Adaptive Regression Splinesover Multiple Linear Regression and Random Forest Regression of rice and wheat and the better algorithmic value of the latter for the Maize Dataset (Apollo Kaneko et al., 2019).

In this paper, the relationship between rainfalls received in the region with the production of the crop by region is identified. Regression techniques such as multiple linear, polynomial linear regression, support vector regression, decision tree regression, random forest regression, and XG- Booster method. Regression method for prediction are preferred over neural networks as regression methods give a lower mean squared error than artificial neural networks. Regression methods with hyper-parameter fine-tuning perform better than neural networks (Khosala

E-mail address: abenny.antony2018@vitstudent.ac.in.

https://doi.org/10.1016/j.envres.2021.111624
Received 23 February 2021; Received in revised form 28 May 2021; Accepted 25 June 2021
Available online 5 July 2021

et al., 2019). This paper differs from the previous research papers in their approach of choosing the relationship of the crop production of specific states with the rainfall received by the state. It almost draws a parallel to the research conducted in Africa. For a pan-Indian view, two states are chosen from that of South India and two states from North India. Two chief crops that have a significant economic contribution to the state are chosen. The crops and the states have high agricultural output and contribute to the high economic output of the country. The methodology that the algorithm discussed in the paper follows can be clearly seen in the way of flowchart in Fig. 1 and as an architecture diagram in Fig. 2.

Section 2 of this paper discusses the literature review of papers that have contributed to developing this paper. Section 3 provides the brief description of the various algorithms used. The results are discussed in Section 4 and the conclusions are drawn in Section 5.

2. Literature review

There has been an improved trend and shift in dynamics towards modern agricultural practices that typically predict what knowledge to expect. Machine learning techniques are a part of data processing and knowledge exploration and focus exclusively on characteristic correlations or patterns among massive datasets or massive relative databases.

1) This paper discusses the crop production rate based on the region's geography, its weather conditions, type of soil found in the region, the composition of the soil, and the type of method used for harvesting the crop. It consists of two types of prediction. One being the traditional statistical method of multiple linear regression problem, the second approach consists of the utilization of a combination of machine learning models such as Artificial Neural Networks, Support Vector Machine, K- Nearest Neighbours, Decision Tree Learning, Random Forest, Gradient Boosting Algorithm and Greedy Forest algorithm. The topological algorithm belonging to the class of Artificial Neural Networks consists of a multi-layered perceptron and back-propagation algorithm for the implementation of the neural network. Support Vector Machine is used to derive a non-linear function using the kernel function. The K-NN method is utilized in increasing the dimensionality of the input function. Decision Tree Learning is used for splitting.
2) This paper discusses the designing of a website for identifying the influence climatic parameters play on certain districts of Madhya Pradesh. The climatic factors taken into account are rainfall, max and min temperature, the approximate transpiration rate, amount of cloud cover. Decision trees were selected as the algorithm for building the model (Khosala et al., 2019). The accuracy of the model used in this paper delivered an accuracy of 76–90 percent.

The general algorithm used in decision tree was:

- Checking for the base cases
- On each attribute a: Normalized information gain by splitting on a
- abest = highest normalized information gain
- Creation of a decision node that splits abest
- Recursively checking the on the sub lists of abest and adding the same to the child nodes

3) This paper deals with the general method of various machine learning models and how the various machine learning models have had a significant impact on agriculture. This paper discusses the various regression, classification, reinforcement methods used in machine learning by keeping in line with the agricultural output. It also discusses how each model behaves in a certain respect to various agricultural components i.e. crop management, soil management, and livestock management. In crop management yield prediction,

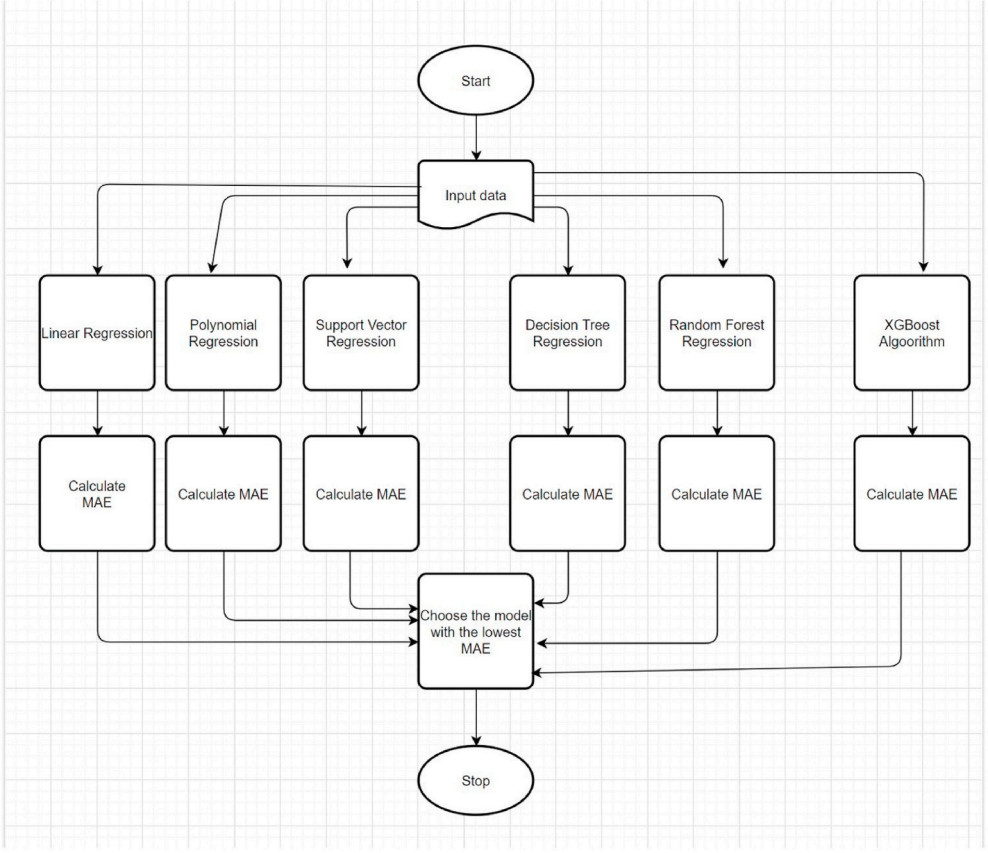

Fig. 1. Flowchart.

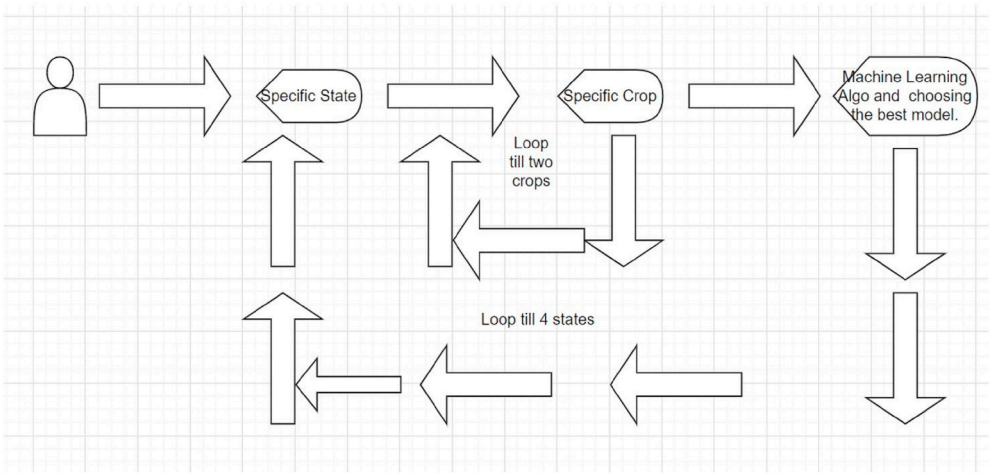

Fig. 2. Architecture diagram.

1) Simple Linear Regression:

disease detection, weed detection, crop quality, and species recognition are discussed. Yield prediction is significant in agriculture and in determining a precision agricultural model. The success of the agricultural output of the crop is highly dependent on the detection of disease that is present in a crop. Machine Learning methods are used to reduce the financial burden of farmers incurred by practicing traditional prevention methods such as pesticides over a large cultivation area using machine learning methods we can specifically target the region that requires pesticides. ML algorithms working with sensors can classify and differentiate crops and weeds and help in segregation. The accuracy in detecting and classifying the quality of the crop is a significant quality and quantity of the crop price and helps in the reduction of waste.

4) This paper deals with the prediction of the production of major Kharif crops in Vishakapattinam. The prediction follows a twofold approach by first determining the monsoon rainfall by the method of the modular artificial neural network, secondly major Kharif crops that can be grown in these rainfall data and with the area by the purpose of support vector regression. Comparison has been done with other machine learning algorithms with this algorithm showing improvement in performance and accuracy. By the application of this technology, crop yields can be increased. The dataset for the place of Vishakapattinam for rainfall was collected from the government website from the years 1901–2015. The agriculture dataset for the crops is collected from the agriculture website of the country. The model was trained by the method of Linear Correlation Analysis from the years 1901–1999. The model is tested on the dataset of rainfall from 2000 to 2017 using one-step lead prediction and finally, rainfall is predicted for the years 2018 and 2019. For yield prediction, important features are selected for feature selection. SVR algorithm was chosen as the best regression algorithm by comparison with other algorithms. The crops analyzed are bajra, maize, rice, and ragi. In this paper, only the crops grown in Vishakapattinam are analyzed and not analyze in a larger geographical area having a significant economic impact.
5) Agriculture plays an important role in sustaining all human activities. Major challenges such as overpopulation, competition for resources pose a threat to the food security of the planet. To tackle the ever-increasing complex problems in agricultural production systems, advancements in smart farming and precision agriculture offer important tools to address agricultural sustainability challenges. Data analytics hold the key to ensure future food security, food safety, and ecological sustainability. Disruptive information and communication technologies such as machine learning, big data analytics, cloud computing, and blockchain can address several problems such as productivity and yield improvement, water conservation, ensuring soil and plant health, and enhance environmental stewardship. The current study presents a systematic review of machine learning (ML) applications in agricultural supply chains (ASCs). Ninety-three research papers were reviewed based on the applications of different ML algorithms in different phases of the ASCs. The study highlights how ASCs can benefit from ML techniques and lead to ASC sustainability. Based on the study findings an ML applications framework for sustainable ASC is proposed. The framework identifies the role of ML algorithms in providing real-time analytic insights for pro-active data-driven decision-making in the ASCs and provides the researchers, practitioners, and policymakers with guidelines on the successful management of ASCs for improved agricultural productivity and sustainability.
6) This paper was developed on the backdrop of winning the 2018 Syngenta Crop Challenge. The model was designed to have superior prediction accuracy with the root mean square being 12 percentage of the average yield. The RMSE reduced even further with the parameters such as perfect weather. The computational results obtained for this model significantly outperformed other popular methods such as Lasso, shallow neural networks (SNN), and regression tree (RT). This paper also proved that environment affected crop yield more than the genotypic makeup of the crop. The factors that were used were the genotypic data and the weather data. The genotypic data were encoded as −1, 0, 1 representing aa, aA, and AA alleles. The weather data is analyzed for a period till 2016. Neural networks are used for weather data the reason being that neural networks can capture the nonlinearities, which exist like weather data, and they learn these nonlinearities from data without requiring the nonlinear model to be specified before the estimation. Two deep neural networks were trained, one for yield and the other for check yield, the differences between the outputs are calculated as the function of prediction of the yield difference. Due to the direct relation of genotype and climatic factors to the yield and check yield. Each neural network has 21 hidden layers and 50 neurons in each layer. After trying deeper network structures, these dimensions were found to provide the best balance between prediction accuracy and limited overfitting. DNN model can approximately preserve the distributional properties of the ground truth yield. Plotted the probability density functions of the ground truth yield and the predicted yield by the DNN model. A major limitation of the proposed model implemented is the black box property, which is shared by many machine learning methods. Although the model captures

Genotype Environment interactions, due to its complex model structure it is hard to produce testable hypotheses that could potentially provide biological insights. To make the model less of a black box, we performed feature selection based on the trained DNN model using the back-propagation method. The feature selection approach successfully found important features and revealed that environmental factors had a greater effect on the crop yield than genotype.

7) This paper was designed to calculate maize data from the maize dataset of Ethiopia, Kenya, Malawi, Nigeria, Tanzania, and Zambia. The remote data input is collected via the MODIS satellite. The raw images are converted to histograms (Pixel Count). This data is used for the prediction of crop yield using LSTM based deep learning model. To improve accuracy Gaussian Process Layer is used. As we analyze data from different countries, each country has its harvest season of maize, UN Food and Agriculture Organization country profiles are used to determine the length and period for each of the country's harvest seasons. Two methods of splits for training are random and chronological. Models are trained for two kinds of splits: random and chronological Due to the varied number of countries incorporated in this research paper the results vary by country. The tests on the last 4 years of countries for Kenya, Tanzania, and Zambia had an average R2 value of 0.50–0.56. The tests performed wore on Ethiopia, Malawi, and Nigeria with an average R2 value of −0.60 to 0.13. Combining all the countries utilizing a combination of all the countries showed a better R2 of 0.63. Randomized splitting of all models achieved a higher level of accuracy. Chronological splits resulted in performance variance due to feature differences, distribution of labels, and quality of data. Results obtained from the Gaussian Processing model emphasize the importance of incorporating spatial features when predicting crop yields.
8) Disease detection in a plant or tree using traditional ways such as the farmer's expert naked eyes is both time and resource consuming and may engender tremendous crop losses. Thus, the early diagnosis and treatment of these diseases can minimize the losses in the whole crop and can improve quality and diversity for the consumer later. With the recent advances in Deep Learning, powerful approaches are developed for both detection and classification that can cope with complex environments. In this paper, efficient deep learning-based architecture for object detection is proposed for the context of Smart Agriculture. The proposed solution combines deep learning and tweaked transfer learning models for object detection with balanced data for every class of images. It can operate in a more complex environment and takes into consideration the state of the input. It aims to automatically detect damages in leaves and fruits, locate them, classify their severity levels, and visualize them by contouring their exact locations. Numerical results reveal that the proposed solution, based on Mask-RCNN achieves higher performances in features extraction and damage detection/localization compared to other pre-trained models such as VGG16 and VGG19.
9) Deep learning has emerged with big data technologies and high-performance computing to create new opportunities for data-intensive science in the multidisciplinary agriculture technologies domain. In this research, a deep learning classification system of diverse plants is presented, to enable precision agriculture applications. This classification problem was achieved thanks to the public dataset "Plant Seedlings Dataset", which contains images of approximately 960 unique plants belonging to 12 species at several growth stages. The database has been from Aarhus University Flakkebjerg Research Station in collaboration with the University of Southern Denmark and Aarhus University. A classification comparison was used to determine which of three pre-trained models; InceptionV3, VGG16, and Xception; reach the best accuracy performance for the database used in this work. Results determined that

- Xception was the best model for plant classification obtaining 86.21 percent, overcoming other networks in 7.37 percent with a time processing around 741 s.
- GPU hardware changes the classification model results impacting strongly in their accuracy score.

3. Methodology

3.1. Dataset

This paper aims to calculate the impact of rainfall affecting the rainfall concerning the crop grown in a certain state. Government acquired data from the meteorological department and the agricultural department of India is used. For rainfall, 'rainfall in India' dataset which is maintained by the Meteorological Department of India is used. This dataset contains the rainfall each district received over a period from 1901 to 2015. The data is given as the rainfall received in each month and it also gives the information of the cumulative rainfall received over the monsoon period present in the country i.e. one period from June to September and another period from October to December. The second dataset 'crop prediction statistics of India from 1997' is used. This dataset contains the information of the important crops grown in each district of each state present in the Indian subcontinent. The crop dataset contains the state name, district name, crop year, season, the crops grown, the area where the crops are cultivated, and the total production of the crop. The two datasets have to be merged based on the districts of the respective states. Only the data after the year 2000 is analyzed. The production and area column are merged and the prediction of the result is to be calculated for Production/Area. This paper discusses the top two crops of two states each from the Southern part and the Northern part of the Indian subcontinent.

3.2. Regression

Regression is a method of modeling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationships between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.

2) Polynomial Regression:

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x. In many types of settings, linear regression may not be able to predict an accurate value.

3) Support Vector Regression:

Support-vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

4) Decision Tree Regression:

A decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The core algorithm for building decision trees called ID3 by J. R. Quinlan employs a top-down, greedy search through the space of possible branches with no backtracking. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain with Standard Deviation Reduction.

5) Random Forest Regression:

Random forest is a flexible, easy-to-use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at the random forest in classification since classification is sometimes considered the building block of machine learning. Random forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

6) XGBoost Algorithm

It implements decision trees that are gradient boosted to improve the speed and its performance. The algorithm was designed to improve the system's efficiency in terms of computing time and memory resources used. The goal for the development is the utilization of the best available resources for training the model. The key implementation features used are:

- Sparse Aware implementation with automatic handling of the missing data values.
- Block Structure to support the parallelized tree construction.
- Continuous training for the further boosting of the fitted model to be implemented on the new data

Regression is a method of modeling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause-and-effect relationships between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

3.3. State and their respective crops

3.3.1. Tamil Nadu

Agriculture constitutes the highest predominant sector of Tamil Nadu's economy, more than 70 percent of the State's population is practicing agriculture or other allied sectors to agriculture as their main profession. The State has an area of 1.3 Lakh sq. km with a gross cropped area of around 63 L.Ha. The Government policy and objectives have been to ensure stability in agricultural production and to increase the agricultural production sustainably to meet the food requirement of a growing population and also to meet the raw material needs of agro-based industries, thereby providing employment opportunities to the rural population. Tamil Nadu has all along been one of the states with a creditable performance in agricultural production with the farmers relatively more responsive and receptive to changing technologies and market forces. The Agriculture Department has taken up the challenge to achieve a higher growth rate in agriculture by implementing several development schemes and also propagation of relevant technologies to step up the production. In- tensive Integrated farming system, massive Wasteland Development Programme, comprehensive watershed development activities, water management through Micro irrigation systems, Organic farming, Soil health improvement through Bio-fertiliser including Green Manuring, adoption of Integrated Nutrient Management (INM), and Integrated Pest Management (IPM) technologies are given priority through various programs, besides crop diversification to fetch better return and value addition to agricultural products are also given priority to improve the economic status of the farming community. Tamil Nadu is considered as one of the states because it contributes a significant factor to the country's agricultural output and it is one of the regions where agriculture is highly dependent on the onset of the monsoon.

3.3.2. Rice

The total cultivable area in Tamil Nadu is 7.99 million hectares and rice cultivation takes up almost 1.93 million hectares. The rice paddy cultivation takes up about 24.16 percent of the cultivable area. The major rice-growing seasons in Tamil Nadu are Jun–Sep, Aug–Jan, and Dec–Apr. The total paddy production is 9.92 million tonnes. The state's contribution to the national output is about 7.8 percentage and it ranks 2nd in all India productivity. The cultivation of rice in Tamil Nadu is affected by water and labor. As water is one of the major constraints of rice cultivation this becomes an important crop for the discussion of the paper.

As rice is one of the major food crops that are highly dependent on rainfall, its analysis is important. The distribution in the production of rice and the rainfall received are shown in Fig. 3. For the prediction of rice output concerning rainfall regression techniques are used for the analysis of data. For analysis, the rice data with its respective rainfall received is separated from the dataset. On this data, there is division into training data and the test data as 80 percent and 20 percent respectively. The model is evaluated on the basis of Mean Absolute error after predictions by Linear Regression, Polynomial Regression Algorithm, Support Vector Regression, Decision Tree Regression, Random Forest Regression, and XGBooster Algorithm.

3.3.3. Bajra

According to Fourth Advance Estimates of the Department of Agriculture and Cooperation, Bajra is cultivated in 7.31 million hectares with a production of 8.06 million tonnes in India during 2015–16, which is 12 percent lesser than the previous year. Bajra is cultivated in 58,000 ha with production was 177,000 tonnes during 2014–15 in Tamil Nadu. Bajra is primarily a food crop and is utilized for about 73.4 and is utilized for many crops. It is also utilized as a feed and fodder crop. In contrast to rice, it is a highly drought-resistant crop and is highly nutrient-rich. The Bajra crops' dataset is separated and analyzed for the prediction of the Bajra production output. The dataset is divided into 80 and 20 proportions for training and testing the data respectively. The Mean Absolute Error is calculated for the Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, and XGBooster Algorithm. The algorithm with the lowest MAE is chosen. The distribution in the production of rice and the rainfall received are shown in Fig. 4.

3.3.4. Kerala

Kerala is a consumer state rather than an effective producer state, it depends on its neighboring states to satisfy its food requirements. Though the state has suitable environmental factors for the cultivation of many diverse crops it has not contributed significantly to the agriculture of India.

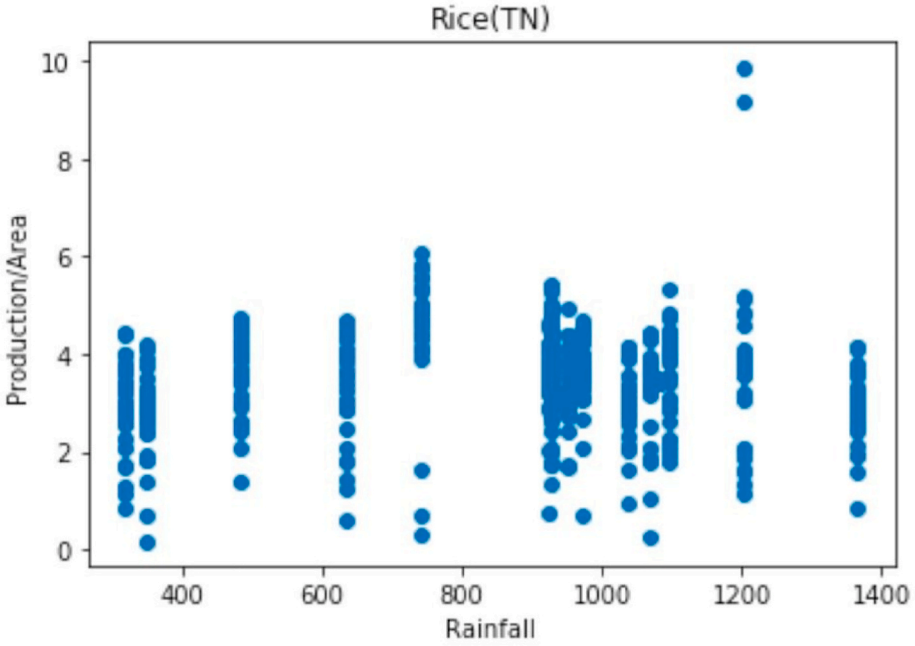

Fig. 3. Rice cultivation and the rainfall received in Tamil Nadu.

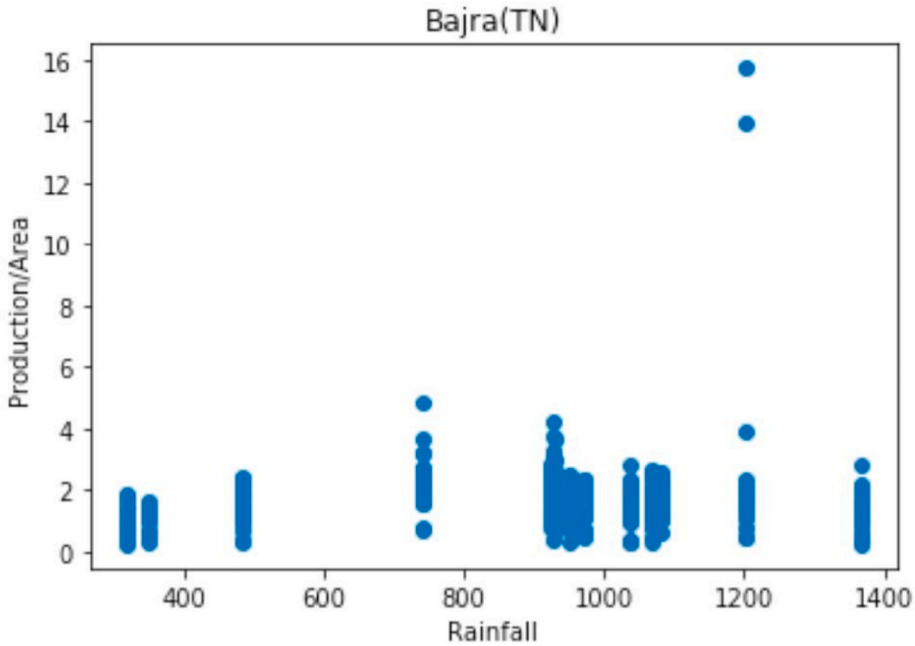

Fig. 4. *B*ajra cultivation and the rainfall received in Tamil.

3.3.5. Coconut

Major coconut growing states in India are Kerala, Tamilnadu, Karnataka, and Andhra Pradesh. Among them, Kerala is the leading state in the area under the cultivation of coconut and its production. Till 1980, Kerala was the major producer of coconut with an 80–85 percent share in national production but later its share in the area under coconut cultivation in the country has fallen sharply from 57 percent in the early 1990s to 43 percent in 2008–09. In Kerala, tender coconut harvesting is very less. It is estimated that less than 2 percent of the total nuts produced are marketed as tender nuts. Har-vesting of matured coconut is a traditional practice in Kerala. Since, copra making, oil extraction, and coir making are principal activities of industrial importance. About 70 percent of matured nuts are converted into copra and out of the total copra produced; about 85 percent is milling copra and 15 percent in the form of edible ball copra. About 30 percent of the nuts are utilized for culinary and other purposes, including dispatches to other States. About 80 percent of the milling copra is converted into oil and the rest along with the ball copra is dispatched to other states. On average a household having 15–20 palm trees is harvesting 1000–1200 nuts per year and getting Rs. 8–10/nut at field level. The trend of registered toddy shop helps in increasing the income level of the farmers having less number of palm trees. They have to register their trees for toddy tapping every year and sell the toddy to registered shops only. The reasons for the declining area under cultivation of coconut are the fragmentation of palm gardens into housing plots and for commercial constructions. In Kerala, the coconut tree is called" Kalpa Vriksham" which essentially means all parts

of a Coconut tree are useful in some way or another. Cocus Nucifera dominates the landscape in many parts, rising to a height of 25 m, and bearing over 50 fruits on average in a year. The trees have many uses; their leaves are used to make sheds, baskets, and doormats, the husk for making coir, the shell for making ladles and spoons, and fruits used for making hair oil or for eating. Co-coconut is a staple ingredient in many Kerala dishes and coconut oil is widely consumed and used to make drinks such as coconut toddy and dishes such as appam. Coconut is also used for making coconut paste which is essential for making traditional curries. The Coconut dataset is separated and analyzed for the prediction of the Coconut production output. The dataset is divided into 80 and 20 proportions for training and testing the data respectively. The Mean Absolute Error is calculated for the Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, and XGBooster Algorithm. The algorithm with the lowest MAE is chosen. The distribution in the production of coconut and the rainfall received are shown in Fig. 5.

3.3.6. Banana

Banana prefers tropical humid lowlands and is grown from the sea level to 1000 m above MSL. It can also be grown at elevations up to 1200 m, but at higher elevations growth is poor. The optimum temperature is 27 °C. Soils with good fertility and an assured supply of moisture are best suited. When the crop is grown as a rain-fed crop it is grown during the April–May season and the irrigated season from August–September. For banana crop cultivation heavy monsoon and severe summer seasons should be avoided for planting. The Mean Absolute Error is analyzed for the algorithms to find the lowest mean absolute error. The distribution in the production of banana and the rainfall received are shown in Fig. 6.

3.3.7. Bihar

Bihar lies in the river plains of the basin of the river Ganga. It is endowed with fertile alluvial soil groundwater resources. This makes

The agriculture of Bihar rich and diverse. Rice, wheat, and maize is the major cereal crop. Arhar, urad, moong, gram, pea, lentils, and khesari are some of the pulses cultivated in Bihar. Bihar is the largest producer of vegetables, which is dominated by potato, onion, eggplant, and cauliflower. In fruit cultivation, it is the largest producer of lychee and the third-largest producer of pineapple, as well as a major producer of mango, banana, and guava. Sugar cane and jute are two other major cash crops of Bihar. The net sown area in Bihar is 60 percent of its geographical area. This percentage is much higher than the all-India average of 42 percent. Such a high percentage of cultivated land is possible for two reasons. First, most of Bihar is a plain area suitable for agriculture. Second, most of the forest had been converted into farmland during the last 2000 years. Currently, land under forest constitutes only 6 percent of the area. North Bihar is a productive agricultural center, while South Bi-har is hindered by its flood and drought-prone geography. In the south, the Ahar-Pyne system of agriculture has long been used to cultivate crops. Rice is cultivated in all districts of Bihar. Autumn rice, Aghani rice, and summer rice are three different varieties of rice grown at three different times of the year. The average production of rice is around 5 million tonnes each year. Some five decades back, wheat cultivation was very restricted to the western districts of Bihar. After the green revolution success, wheat was planted by Bihari farmers on a larger scale, and wheat now occupies the status of major crop of the rabi (spring) season. The average annual wheat production is approximately 4–4.5 million tonnes. Maize is also cultivated, with an average annual production level of approximately 1.5 million tonnes and a steady positive trend in production. The leading producer districts are Khagaria and Saharsa. Pulses such as moong, arhar, peas, and khesari are grown, more in southern than in northern Bihar. The leading districts are Patna, Bhojpur, Aurangabad, and Nalanda. The total area under vegetable cultivation is currently about 11 percent of the state's gross sown area and is increasing. The important vegetable crops include potato, onion, tomato, cauliflower, and brinjal. Hajipur in Vaishali is famous for an early variety of cauliflower that reaches the market in the last week of September. Production of vegetables is well dispersed over the districts, with a concentration of production in some particular districts. Apart from Patna and Nalanda (Jehanabad), where vegetable production is quite extensive, the other districts with high shares in total vegetable production are Vaishali, Muzaffarpur, West Champaran, East Champaran, Katihar, and Begusarai.

3.3.8. Wheat

Wheat is grown on more land area than any other food crop (220.4 million hectares, 2014). World trade in wheat is greater than for all other crops combined. In 2017, world production of wheat was 772 million tonnes, with a forecast of 2019 production at 766 million tonnes,

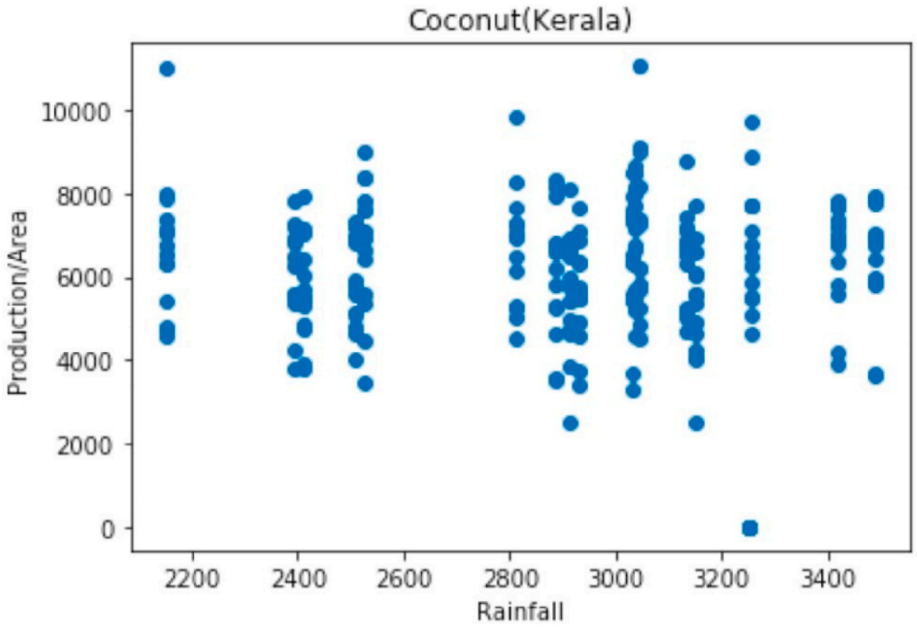

Fig. 5. Coconut cultivation and the rainfall received in Kerala.

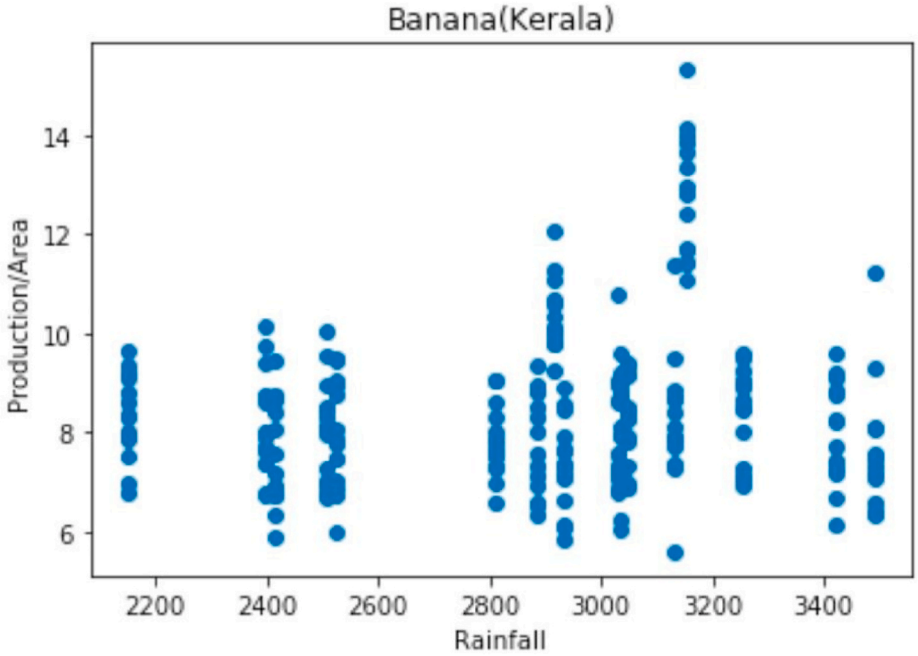

Fig. 6. Banana cultivation and the rainfall received in Kerala.

making it the second most-produced cereal after maize. Wheat is a crop of temperate climate. The ideal temperature for its cultivation is about 15°–20 °C and requires a moderate amount of rainfall of 25–75 cms. It can be grown in drier areas with the help of irrigation. The mean absolute error is analyzed for the algorithms and the lowest is chosen as the one that is suitable for prediction. The distribution in the production of wheat and the rainfall received are shown in Fig. 7.

3.3.9. *Maize*

Bihar accounts for over 80 percent of India's 6–7 million tonnes (mt) annual production of rabi maize, which is sown in November–December and marketed in May–June. Maize prices at Gulabbagh, the state's largest mandi for the feed grain in Purnia, are currently averaging Rs 1250 per quintal. This is way below the Rs 2000–2400 rates in wholesale markets across India from last June–July till January and also the government's minimum support price (MSP) of Rs 1760/quintal for 2019–20. Maize is a major crop in Bihar's eastern districts north of the Ganga. The mean absolute error is analyzed for the algorithms and the lowest is chosen as the one that is suitable for prediction. The distribution in the production of maize and the rainfall received are shown in Fig. 8.

3.3.10. *Punjab*

Punjab is one of the most fertile regions on Earth. The region is ideal for growing wheat, rice, sugarcane, fruits, and vegetables. Punjab is called the "Granary of India" or India's bread-basket. It produces 20

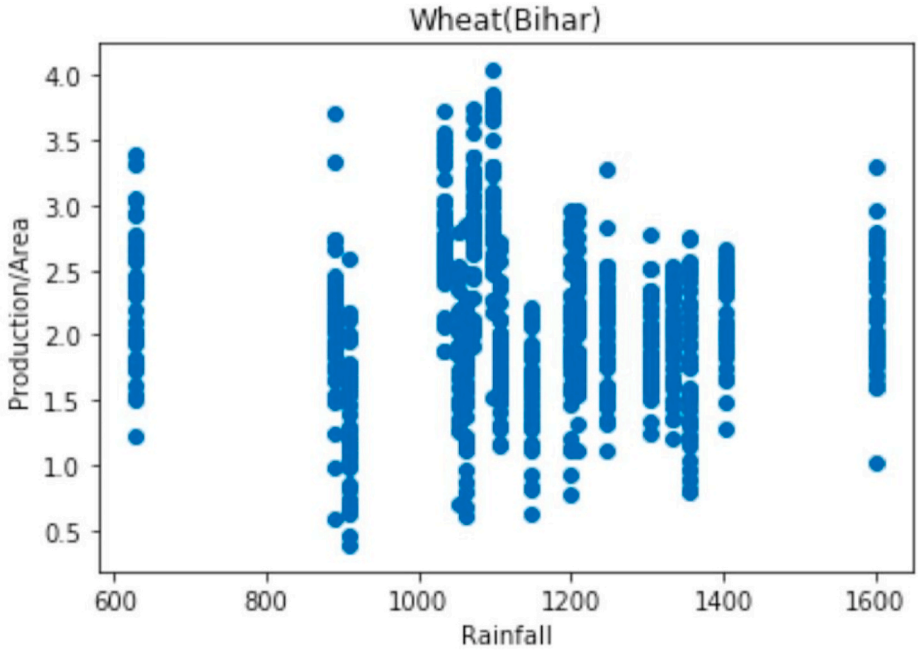

Fig. 7. Wheat cultivation and the rainfall received in Bihar.

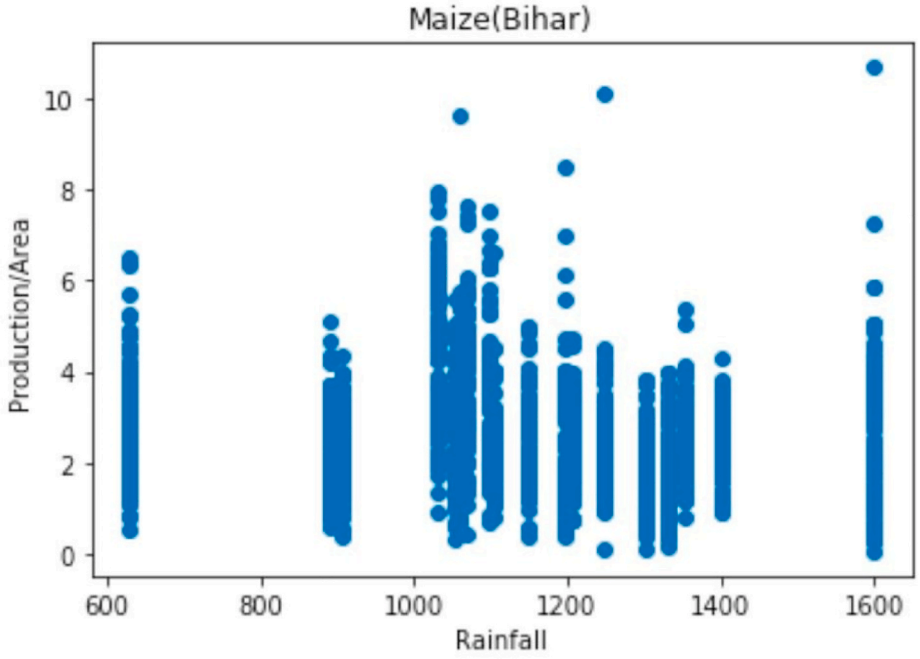

Fig. 8. Maize cultivation and the rainfall received in Bihar.

percent of India's wheat and 9 percent of India's rice. On a global scale, this represents 3 percent of the world's production of these crops, so the Indian Punjab produces 2 percent of the world's cotton, 2 percent of its wheat, and 1 percent of the world's rice. The largest grown crop is wheat; however, other important crops are rice, cotton, sugarcane, pearl millet, maize, barley, and fruits.

3.3.11. Wheat

The crops dataset of the Punjab wheat dataset is separated and analyzed using the Mean Absolute Error of different prediction algorithms. Wheat has a different distribution factor as it is not highly dependent on water as rice. The distribution in the production of wheat and the rainfall received are shown in Fig. 9.

3.3.12. Rice

The crops dataset of the Punjab rice dataset is separated and analyzed using the Mean Absolute Error of different prediction algorithms. Rice is highly dependent on water. The distribution in the production of rice and the rainfall received are shown in Fig. 10.

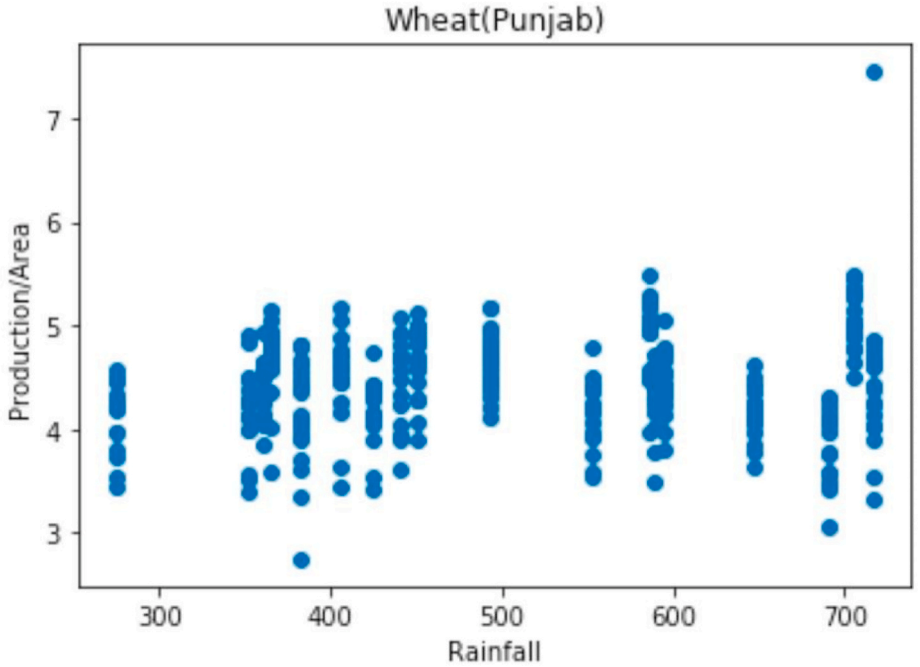

Fig. 9. Wheat cultivation and the rainfall received in Punjab.

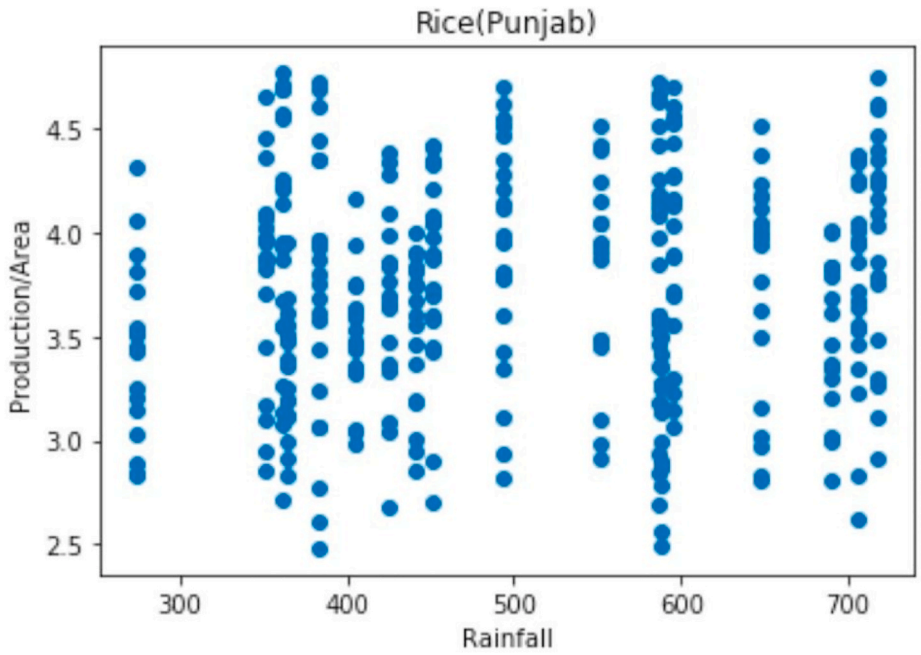

Fig. 10. Rice cultivation and the rainfall received in Punjab.

4. Result

4.1. Tamil Nadu

By observing the state's rainfall we can deduce that it receives 48 percent from the North-East monsoon, and 32 percent from the South-West monsoon.

4.2. Rice

The state has reached its peak production capacity at 1200 cm of rainfall. It has had a highly consistent period of production when it has received 900–1000 cm of rainfall. For the rice dataset, the Support Vector Regression has the lowest MAE of 0.659. Hence it is chosen as the suitable algorithm. The difference of the MAE can be viewed in Table 1 and graphically visualized in Fig. 11. Due to the detailed availability of data for rice as it is a perennial crop in Tamil Nadu for all the respective years all the regression algorithm have almost the same Mean Absolute Error.

4.3. Bajra

The state has reached its production peak at a capacity of 1200. It has its highest consistency from 900 to 1000 cm of rainfall. For this dataset, the Random Forest Regression has the least MAE of 0.600. It is taken as the prediction algorithm. The difference of the MAE can be viewed in Table 2 and graphically visualized in Fig. 12. Due to Bajra not being a perennial crop in the state of Tamil Nadu the difference between the mean of the Mean Absolute Error and its median is significant.

Table 1
Results for rice in Tamil Nadu.

Algorithm	Value
Linear regression	0.7401
Polynomial regression	0.6972
Support vector regression	0.6539
Decision tree regression	0.6940
Random forest regression	0.6844
XGBoost algorithm	0.6931

4.4. Kerala

Meanwhile, its extreme eastern fringes experience a drier tropical wet and dry climate. Kerala receives an average annual rainfall of 3107 mm – some 7030 crore m3 of water. This compares to the all-India average is 1197 mm. Parts of Kerala's lowlands may average only 1250 mm annually.

4.5. Coconut

The state has reached its production peak at 2200 and 3100 cm of rainfall. It is highly consistent production at 2800–2900 cm of rainfall. Random Forest Regression delivers the lowest MAE of 1056.54 and is taken as the prediction algorithm. The difference of different algorithms of the MAE can be viewed in Table 3 and graphically visualized in Fig. 13.

4.6. Banana

Peak production capacity is achieved at 3190 to 3195 and high consistency is achieved at 2400, 2500, and 3100. Has the lowest MAE at XGBoostAlgorithm with a value of 0.793. The difference of different algorithms of the MAE can be viewed in Table 4 and graphically visualized in Fig. 14.

4.7. Bihar

The monsoon season in Bihar is usually becoming very unpredictable and erratic. In the last decade, Bihar recorded 6 to 7 drought years. The last time Bihar truly had a wet year was in 2007. From 2009 till 2019 Bihar experienced erratic monsoons with least rainy days. The concern is growing because due to rapid industrialization and cutting of trees led to serious climate change in the state. Begusarai district of Bihar is one of the most affected areas due to climate change. Begusarai used to receive around 1200–1300 mm of annual rainfall annually with 1750 mm in 2007 but now it receives less than 500–600 mm due to which there's less rice cultivation and other crops which require high rainfall. This area of Bihar was known to grow varieties of rice but farmers are not willing to grow rice due to a shortage of rain and irrigation. There's a concern

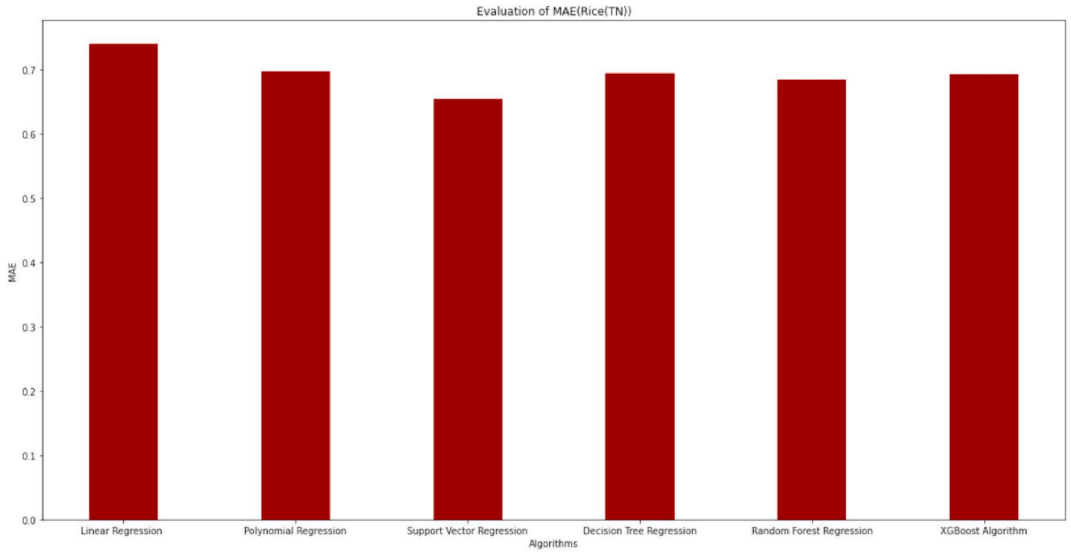

Fig. 11. Mean Absolute Error of different algorithms for Rice in Tamil Nadu.

Table 2
Results for bajra in Tamil Nadu.

Algorithm	Value
Linear regression	0.7254
Polynomial regression	0.6390
Support vector regression	0.6189
Decision tree regression	0.6163
Random forest regression	0.6008
XGBoost algorithm	0.6170

about climate change that is serious and fatal because now people are experiencing a shortage of groundwater. Soon after Mid June this the rainy season commences and continues till the end of September, the beginning of this season occurs when a storm from the Bay of Bengal passes over Bihar. The commencement of monsoon may be as early as the last week of May or as the first or second week of July. The rainy season begins in June. The rainiest months are July and August. The rains are the gifts of the southwest monsoon. There are three distinct areas where rainfall exceeds 1800 mm. Two of them lie on the northern and north-western wings of the state and the third lies in the Netarhat pat. The southwest monsoon normally withdraws from Bihar in the first week of October.

4.8. Wheat

The state achieves the highest production at 1100 cm in rainfall and consistent at 1000–1200 cm in rainfall. Has the lowest MAE at 0.3595 at XGBoost Algorithm. The difference of different algorithms of the MAE

Table 3
Results for coconut in Kerala.

Algorithm	Value
Linear regression	1363.595
Polynomial regression	1361.503
Support vector regression	1401.708
Decision tree regression	1058.264
Random forest regression	1056.543
XGBoost algorithm	1060.118

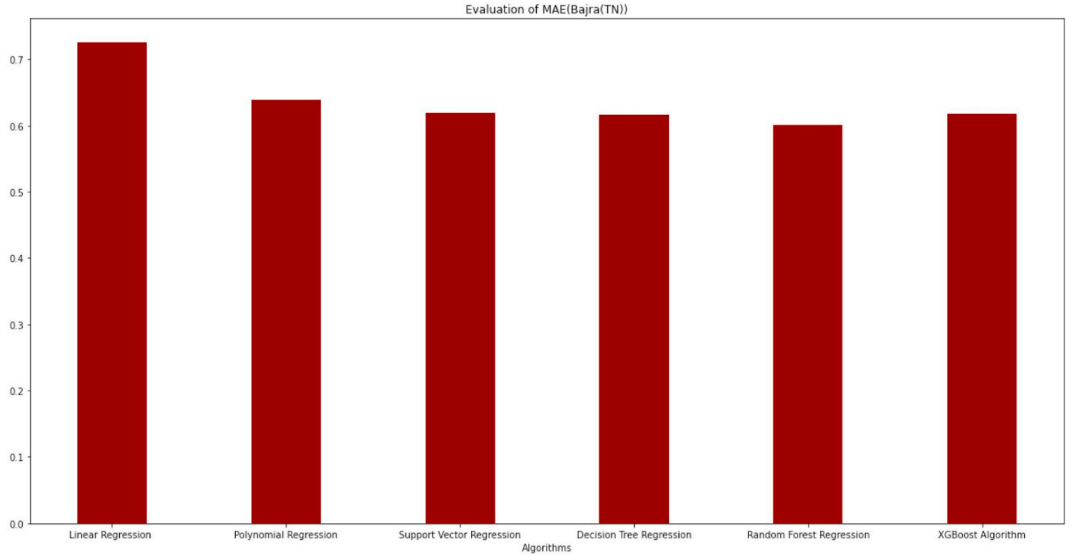

Fig. 12. Mean Absolute Error of different algorithms for Bajra in Tamil Nadu.

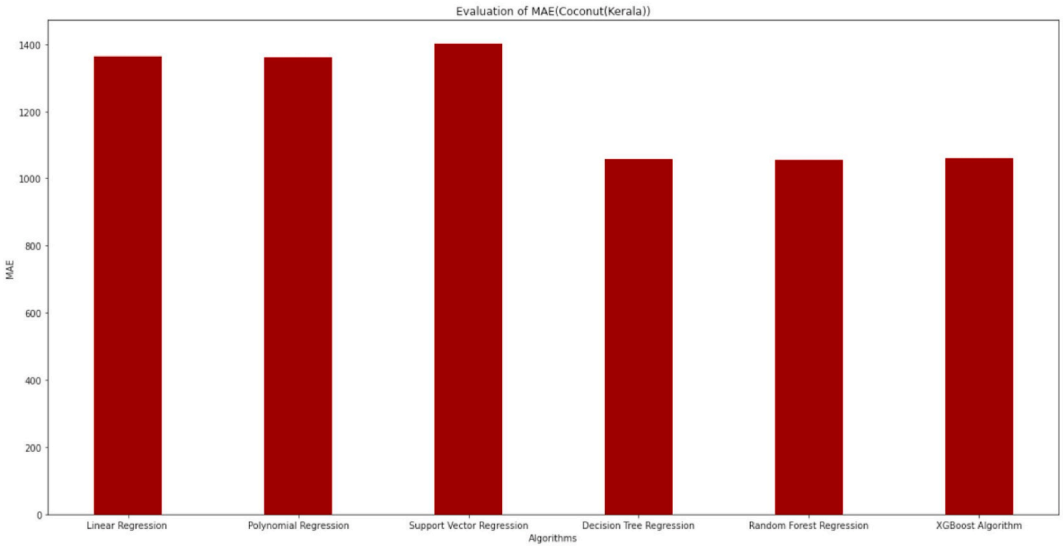

Fig. 13. Mean Absolute Error of different algorithms for Coconut in Kerala.

Table 4
Results for banana in Kerala.

Algorithm	Value
Linear regression	1.0902
Polynomial regression	1.0653
Support vector regression	1.0789
Decision tree regression	0.8014
Random forest regression	0.8211
XGBoost algorithm	0.7937

can be viewed in Table 5 and graphically visualized in Fig. 15.

4.9. Maize

The state achieves the highest production at 1600 cm in rainfall and consistent at 1000–1100 cm in rainfall. Has the lowest MAE at Decision Tree Regression with the value of 0.881. The difference of different algorithms of the MAE can be viewed in Table 6 and graphically visualized in Fig. 16.

4.10. Punjab

Monsoon season provides most of the rainfall for the region. Punjab receives rainfall from the monsoon current of the Bay of Bengal. This monsoon current enters the state from the southeast in the first week of July. The winter season remains very cool with temperatures falling below freezing at some places. Winter also brings in some western disturbances. Rainfall in the winter provides relief to the farmers as some of the winter crops in the region of Shivalik Hills are entirely dependent on this rainfall. As per meteorological statistics, the sub-Shivalik area receives more than 100 mm (3.9 in) of rainfall in the winter months.

Table 5
Results for wheat in Bihar.

Algorithm	Value
Linear regression	0.5011
Polynomial regression	0.4888
Support vector regression	2.2023
Decision tree regression	0.3603
Random forest regression	0.3634
XGBoost Algorithm	0.3595

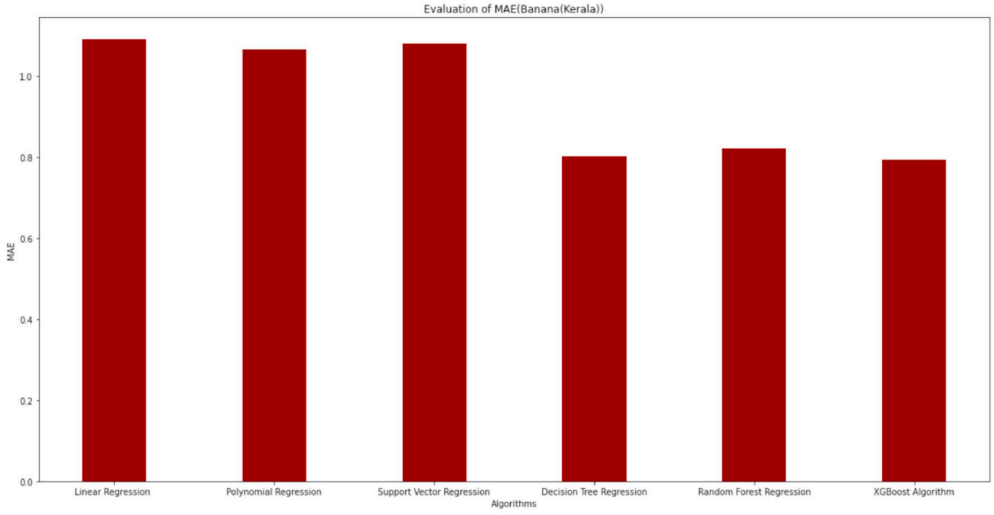

Fig. 14. Mean Absolute Error of different algorithms for Banana in Kerala.

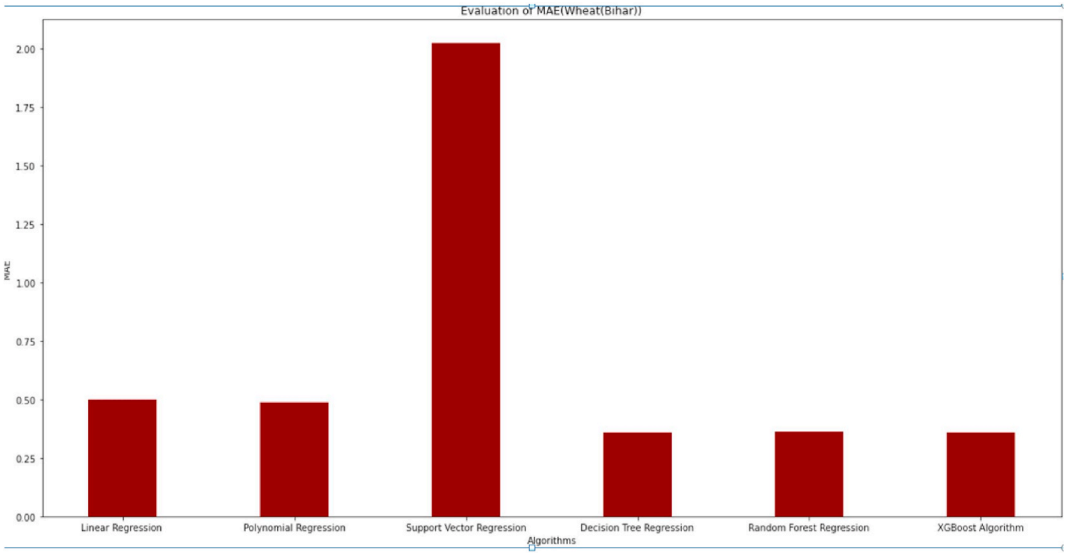

Fig. 15. Mean Absolute Error of different algorithms for wheat in Bihar.

Table 6
Results for maize in Bihar.

Algorithm	Value
Linear regression	0.9961
Polynomial regression	0.9542
Support vector regression	0.9341
Decision tree regression	0.8817
Random forest regression	0.8818
XGBoost Algorithm	0.8822

4.11. Wheat

Has the highest production at 700 cm at rainfall and 350–450 cm of rainfall has the most consistent production of rainfall. Has the lowest MAE at Random Forest Regression with the value of 0.36120. The difference of different algorithms of the MAE can be viewed in Table 7 and graphically visualized in Fig. 17.

4.12. Rice

Has the highest production at 370 and 710 cm of rainfall and 590–600 cm of rainfall gives the consistent value of rainfall. The lowest MAE is at 0.349. The difference of different algorithms of the MAE can be viewed in Table 8 and graphically visualized in Fig. 18.

Table 7
Results for wheat in Punjab.

Algorithm	Value
Linear regression	0.4525
Polynomial regression	0.4364
Support vector regression	0.4274
Decision tree regression	0.3630
Random forest regression	0.3612
XGBoost algorithm	0.3618

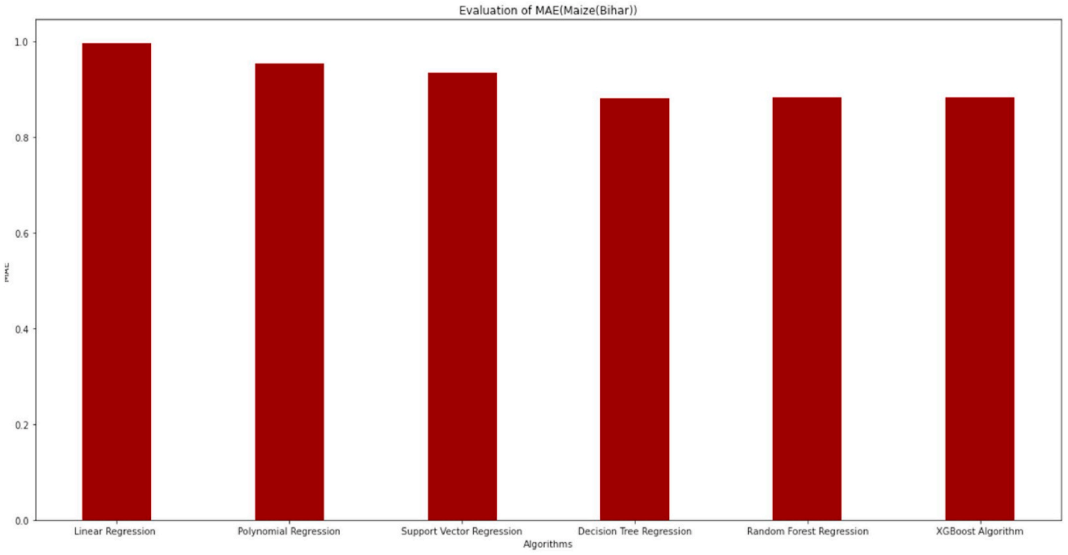

Fig. 16. Mean Absolute Error of different algorithms for Maize in Bihar.

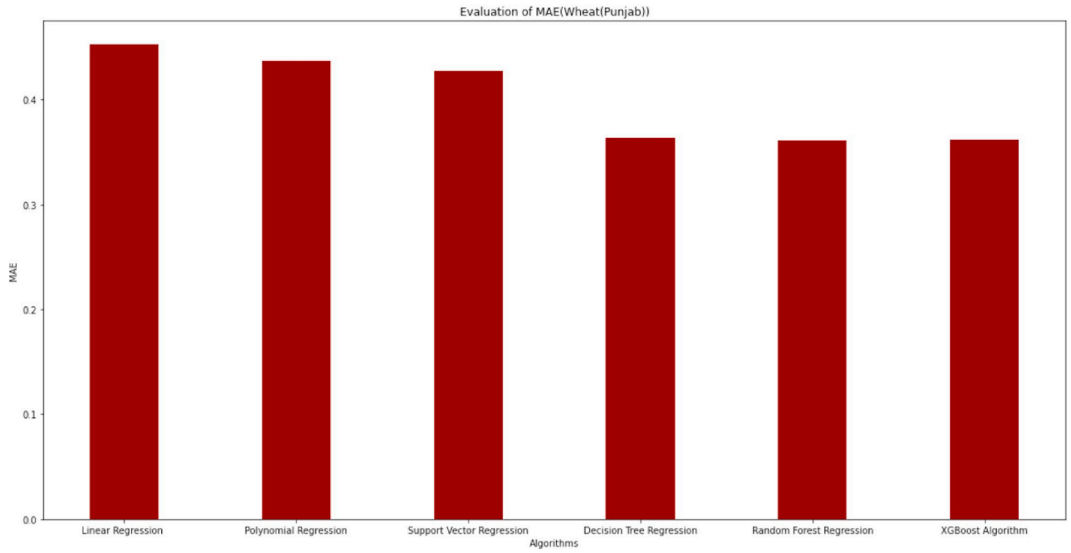

Fig. 17. Mean Absolute Error of different algorithms for Wheat in Punjab.

Table 8
Results for rice in Punjab.

Algorithm	Value
Linear regression	0.4216
Polynomial regression	0.4150
Support vector regression	0.4191
Decision tree regression	0.3491
Random forest regression	0.3509
XGBoost algorithm	0.3507

5. Conclusion

Two crops belonging to four states each are analyzed. For diversity and all Indian inclusiveness, two states from the Southern part of India and two states from the northern part of India are chosen. These crops and states are also chosen for their consumption rate and their contribution to the general economy of the country. This paper aims to help the farmers in improving their farming practices by using scientific, statistical, and data science methodologies. Each crop had a specific optimum algorithm based on the least Mean Absolute Error. Rice for Tamil Nadu and rice in Punjab had two different algorithms. Mean Absolute Error was taken as the statistical evaluator for the models This work can be further extended to all the crops grown in India and done for all the states in India. In farming, though rainfall is also one of the most important factors, it also depends on the soil quality, irrigation facility, and a multitude of other factors. Using these factors future machine learning problems can be constructed. . As is evident from the results each crop in each state has each algorithm as the optimum algorithm. As more modern regression methods are developed and farming practices improved we will be able to get more accurate predictions.

Author contribution

Benny Antony: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Software, Validation, Writing – review & editing.

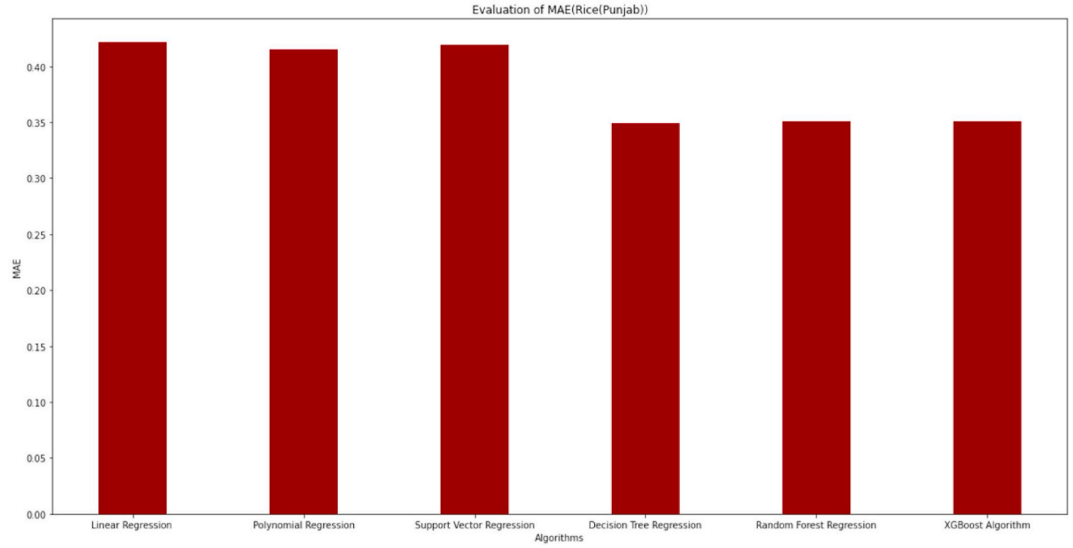

Fig. 18. Mean absolute error of different algorithms for Rice in Punjab.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Kaneko, Apollo, Kennedy, Thomas, Mei, Lantao, Sintek, Christina, Burke, Marshall, Ermon, Stefano, Lobell, David, 2019. Deep Learning for Crop Yield Prediction in Africa.

Boukhris, Louay, Abderrazak, Jihene Ben, Besbes, Hichem, 2020. Tailored deep learning based architecture for smart agriculture. 2020 Int. Wireless Commun. Mobile Comput. (IWCMC).

Khakhi, Saeed, Wang, Lizhi, 2019. Crop yield prediction using deep neural networks. Front. Plant Sci.

Khosala, Ekaansh, Dharavath, Ramesh, Priya Sharma, Rashmi, 2019. Crop yield prediction using aggregated rainfall based modular artificial neural networks and support vector regression. Environ. Dev. Sustain.