National Institute of Technology, Patna

JAN 2022 - MAY 2022

The Crop Yield Prediction Using Machine Learning

By

Priyanshu Bhardwaj (1906023)
Pranjal Mishra (1906024)
Khot Kshitij Dadaso(1906086)

Under the Supervision of

Dr Anushree Tripathi
Department of Computer Science and Engineering
NIT PATNA

CERTIFICATE

This is to certify that **Pranjal Mishra(1906024), Kshitij Khot Dadaso(1906086), Priyanshu Bhardwaj(1906023)** have carried out the Minor Project-1 (CS6490) entitled as "**The Crop Yield Prediction using Machine Learning**" during their 6th semester under the supervision of Dr. Anushree Tripathi, Department of Computer Science and Engineering in partial fulfillment of the requirements for the award of **Bachelor of Technology** degree in the **Computer Science and Engineering**, **National Institute of Technology Patna.**

Dr. Anushree Tripathi
CSE Department
NIT PATNA

Dr. J.P. Singh
Head of Department
CSE Department
NIT PATNA

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

DECLARATION

We students of the 6th semester hereby declare that this project entitled "**The Crop Yield Prediction using Machine Learning**" has been carried out by us in the Department of Computer Science and Engineering of National Institute of Technology Patna under the guidance of Dr. Anushree Tripathi, Department of Computer Science and Engineering, NIT Patna. No part of this project has been submitted for the award of a degree or diploma to any other Institute.

Name	**Signature**
Priyansh Bhardwaj(1906023)	----------
Pranjal Mishra (1906024)	----------
Kshitij Khot (1906072)	----------

CONTENTS

Acknowledgment

We would like to express our special thanks to our project supervisor Dr. Anushree Tripathi who gave us the golden opportunity to do this wonderful project on the topic The Crop yield Prediction using Machine Learning , which helped us in doing a lot of research work and we came to know about so many things which were new to us. We are really thankful to her.
Secondly, we would also like to thank all those people who helped us a lot in finalizing this project within the limited time frame.

Date:29th April 2022

Group-Members:

Priyanshu Bhardwaj	1906023
Pranjal Mishra	1906024
Kshitij Khot	1906086

Abstract

In India, we all know that Agriculture is the backbone of the country. Agriculture is the first and foremost factor which is important for survival. Machine learning (ML) could be a crucial perspective for acquiring real-world and operative solutions for crop yield issues.

Considering the present system including manual counting, climate smart pest management and satellite imagery, the results obtained aren't really accurate.

The project focuses on predicting the crop yield in advance by analyzing factors like district state, season, crop type using various supervised machine learning techniques. This helps the farmers to know the crop yield in advance to plan and choose a crop that would give a better yield.

Keywords - agriculture; machine_learning; crop_yield_prediction; logistic_regression; naïve bayes; random forest; weather_api;

INTRODUCTION

Agriculture, since its invention and inception, has been 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 an important sector for the Indian economy and also for the human future. It also contributes an outsized portion of employment. Because as the time passed the requirement for production has increased exponentially. So as to produce in mass quantities 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 crops. 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 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 analyzed 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 So as to perform accurate prediction and stand on the various machine learning classifiers like Linear Regression, Decision tree, Random Forest etc. are applied to urge a pattern. By applying the above machine learning classifiers, we came to the conclusion that the Random Forest algorithm provides the foremost accurate value. System predicts crop prediction from the gathering of past data. Using past information on weather, temperature and a number of other factors the information is given. The Application which we developed, runs the algorithm and shows the list of crops suitable for entering data with predicted yield value.
India being a developing nation with a rapidly growing population, food consumption is high and will increment in the near future. Thus to establish food security, vertical advancement in agribusiness is of great importance.. Likewise, it turns out to be fundamental to replicate and forecast the crop yield under encompassing conditions, prior to the application phase for efficient crop management. Since the association between crop yield and climatic and non-climatic parameters are nonlinear and incorporate some difficulties, machine learning may result in a rewarding option for crop yield prediction.

Machine learning administers specific strategies to characterize principles and patterns in extensive datasets relevant to crop yield with eminent forecasting capacity. Also, it would be able to improvise the forecasting model. At present agricultural advances are largely redirected to machine learning algorithms since it has augmented crop yield with limiting information cost. Machine learning algorithms enhance the agriculturists to improve the crop selection and crop yield forecast.

Now we are going to focus on the practical implementation of the solution.

Objective

To bridge the gap between technology and agriculture sector:

- By developing a predictive model using machine learning techniques
- By proposing a user-friendly Web Interface to interact with the model
- Encouraging farmers to get into the age of information.

Motivation

In India, we all know that Agriculture is the backbone of the country.

Agriculture is the driving catalyst for the rural economy. It is not only an enormous aspect of a growing economy but it is essential for us to survive. It is anticipated that India's food demand is going to rise between 60% and 95% by 2050. However, Current Agricultural Techniques have some flaws.So there is need of systematic use of technology for the efficient management and yield of agricultural products.

Currently, the problem with the Indian farmers is that most of them are unaware of the fact that which crop can better grow in their farm because of lack of information or wrong information. New sorts of hybrid varieties are produced day by day, but they are not aware which is gonna give more yield. The unnatural techniques spoil the soil by creating an imbalance of nutrients. If some crops fail there is a huge loss on farmers' heads. This is not good for the long term.

For the resolution of this problem, we are proposing a website where the user (or farmer) can enter the data related to the farming area such as area of land, location of the farm, etc and get the prediction of crop yield or the suggestion about which crop will be the most suitable for the user. Due to the accurate information about the crop yield the loss factor over the farmer will decrease. It will increase the diversity of food across the country and help in improving the soil quality too.

Literature survey

In many of the research papers it has been found that everyone uses climatic factors like rainfall, sunlight and agricultural factors like soil type, nutrients possessed by the soil (Nitrogen, Potassium, etc.) but the problem is we need to gather the data and then a third party does this prediction and then it is explained to the farmer and this takes a lot of effort for the farmer and he doesn't understand the science behind these factors. To make it simple and which can be directly used by the farmer, this paper uses simple factors like which state and district the farmer is from, which crop and in what season (as in Kharif, Rabi, etc.)[1].

[2] In this paper, a new hybrid regression-based algorithm, Reinforcement Random Forest, is proposed which displays significantly enhanced performance over traditional machine learning techniques like the random forest, decision tree, gradient boosting, artificial neural network and deep Q-learning. The new strategy executes reinforcement learning at every selection of a splitting attribute amid the process of tree construction for the efficient utilization of the available samples. They analyze the variable significance measure to select the most substantial variable for node splitting process in the model development and promote efficient utilization of training data.

[3] This paper proposes a hybrid deep learning-based crop yield prediction system using deep belief networks (DBN) and fuzzy neural networks systems (FNN). DBN is a combination of statistics and probability with neural networks. Though DBN performs better for nonlinear systems, the algorithm alone cannot provide satisfactory results in terms of robustness, model accuracy and learning speed, which is predominantly due to gradient diffusion.The proposed model efficiently predicts the results outperforming the other models by preserving the original data distribution with an accuracy of 92%.

[4] The reviews done have shown good inclination towards hybrid models and deep learning techniques as means of crop yield prediction. The study also reviewed the works done by researchers in assessing the influence of various factors on crop yields and temperature and precipitation have been found to have maximum influence on the yields of different crops. Apart from climatic factors, agronomic practices adopted by farmers at various stages of growth of a plant also have considerable influence of the final yield of crop.

[5] People have conducted experiments on Indian government dataset and it's been established that the Random Forest machine learning algorithm gives the best yield prediction accuracy. Results reveal that Random Forest is the best classifier when the relevant parameters are combined.

[6] People have implemented crop yield prediction by using only the random forest classifier. Various features like rainfall, temperature and season were taken into account to predict the crop yield. Other machine learning algorithms were not applied to the datasets. With the absence of other algorithms, comparison and quantification were missing thus unable to provide the apt algorithm.

[10] In this approach, crop yield data for several years is considered and the set of parameters most effective or contributing to the yield variations are determined. Accepting these efficacious parameters as independent and harvest yield as dependent variables, empirical equations are formulated to compute the coefficients of these parameters. These coefficients are used to estimate the final crop yields. Every

statistical model determines one set of parameters. Such techniques are relatively less expensive and easy in application and also they do not need any prior information on the various physiological processes involved in the growth of the plant or predefined structure of the model.

The variations of the same technique in different studies is owing to the use of different crops and diverse parameters of study for the models. Apart from weather and soil parameters, agronomic practices adopted by farmers have also shown considerable effect on the final yield of crops. The evidence shows that there is no standardization of parameters and how the parameters are being tuned. The focus is required to study the optimized deduction of parameters on which the model is based.

Advanced techniques of machine learning like deep learning have shown good potential in dealing with huge amounts of data. Also their efficiency in learning through pattern recognition in the data without any outside training has made them particularly suitable for the area of crop yield predictions.

[9] This paper focuses on the prediction of crop and calculation of its yield with the help of machine learning techniques. Several machine learning methodologies used for the calculation of accuracy. Random Forest classifier was used for the crop prediction for the chosen district. Implemented a system to crop prediction from the collection of past data. The proposed technique helps farmers in decision making of which crop to cultivate in the field.

Proposed Model

Through The crop yield prediction system better planning and decisions can be made by the farmer before sowing the seeds. This predictive Modal has been built in the following steps:

1. Preparing Data / Data PreProcessing
2. Selection of the algorithm having precise results
3. Training the Model using the selected Algorithm
4. System Analysis

Data Preprocessing and Feature selection

The dataset collected from the Indian Government Repository is first cleaned as It may contain some incomplete, redundant, inconsistent data. Therefore in this step such redundant data should be filtered. Data should be normalized. After that the irrelevant data which is not a feature for our model should be dropped and finally we are left with the feature selected dataset. This includes factors affecting yield and production.

The dataset contains the tuples having the following column headers:

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

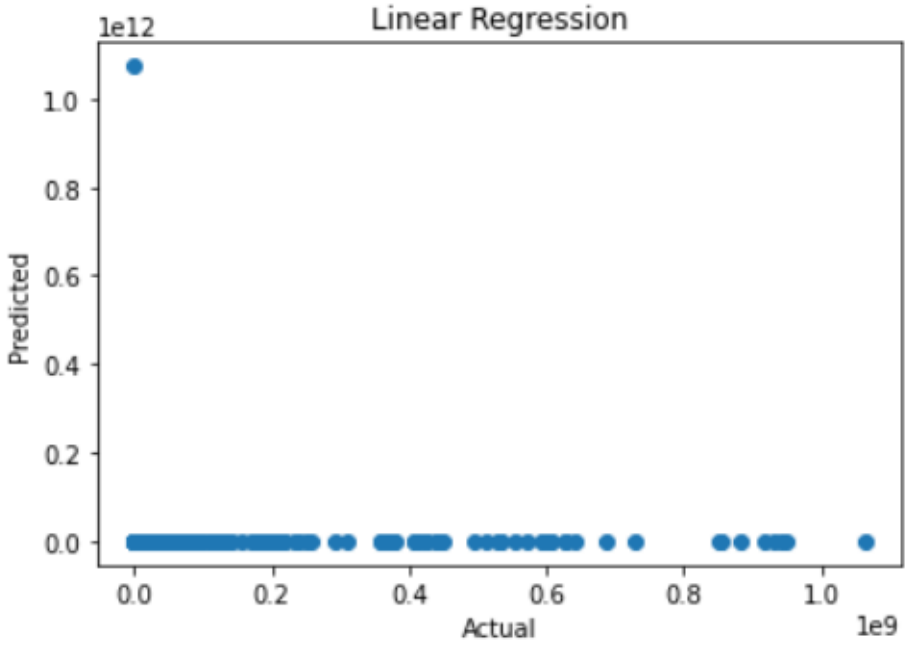
	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0
...	...	...	...	...	...	...	...

Visualizing Dataset

The preprocessed data consists of all the fields left after dropping the state_Name column additionally we added a new column named Yield which indicates Production per unit Area.

Algorithm Selection

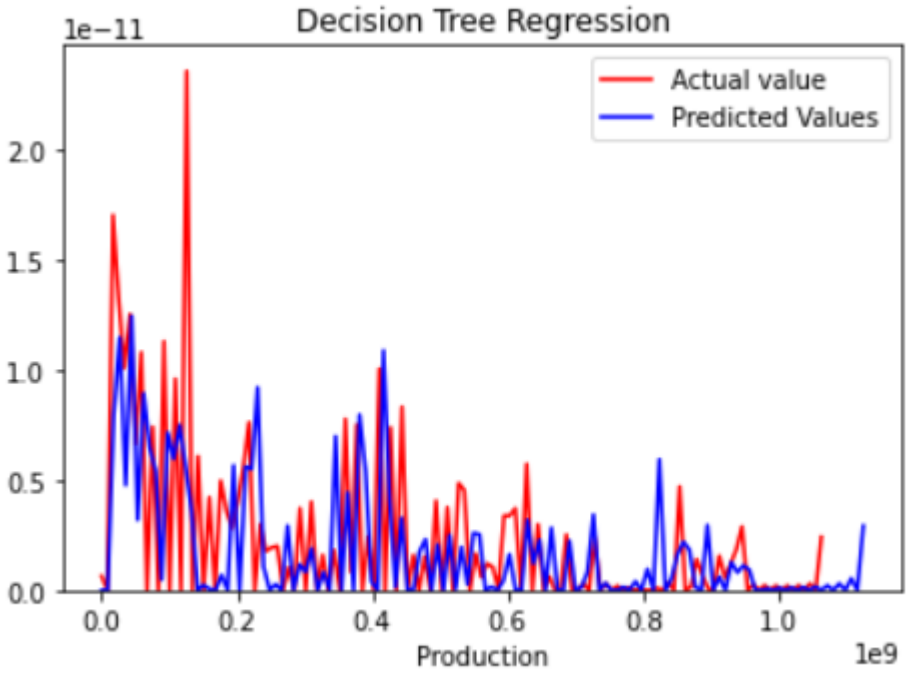
Linear Regression - Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis. It takes into account the significance of independent variables and predicts the dependent variable.



Linear regression

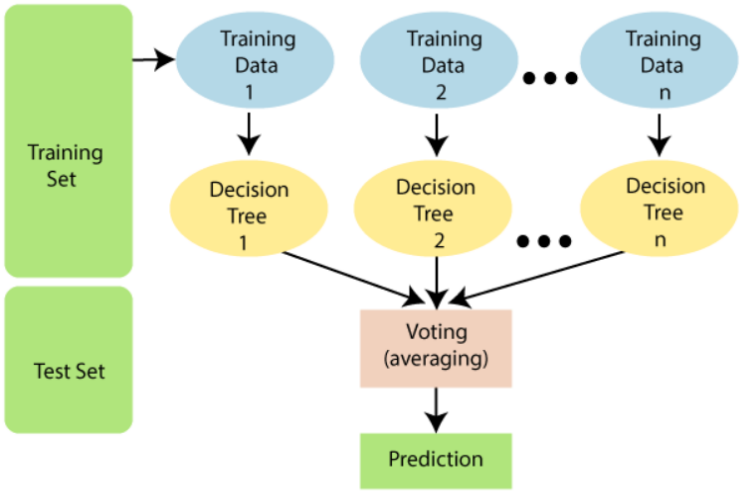
It can be clearly seen that the linear regression is not suitable for this dataset.

Decision Tree - Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

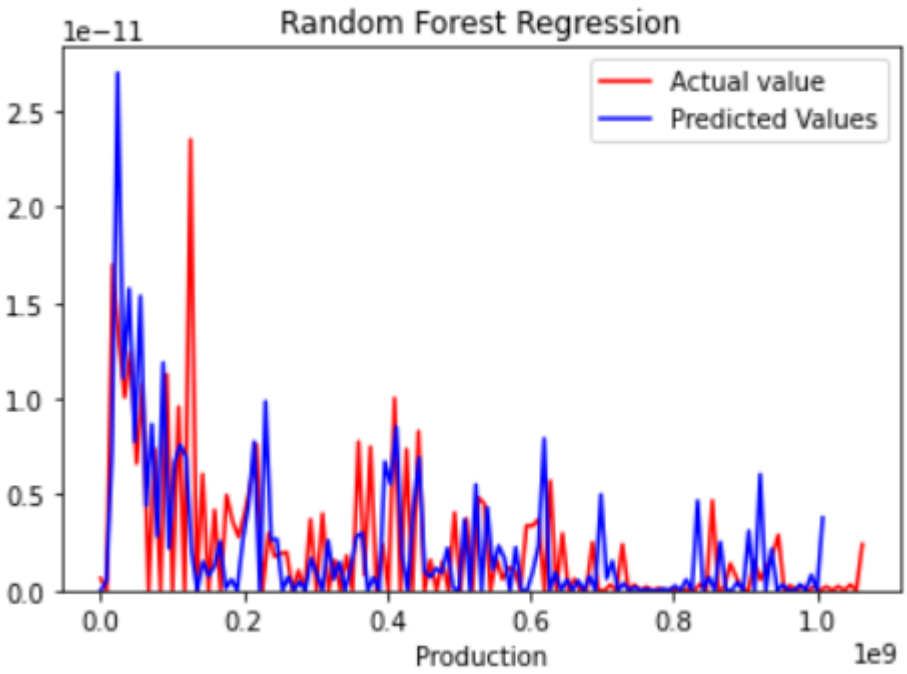


Decision tree regression

Random Forest - Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.



Random forest Algorithm (source: Javatpoint.com)



Random Forest regression

There are a number of key advantages and challenges that the random forest algorithm presents when used for classification or regression problems. Some of them include:

- Reduced Risk of overfitting
- Provides flexibility
- Easy to determine feature importance

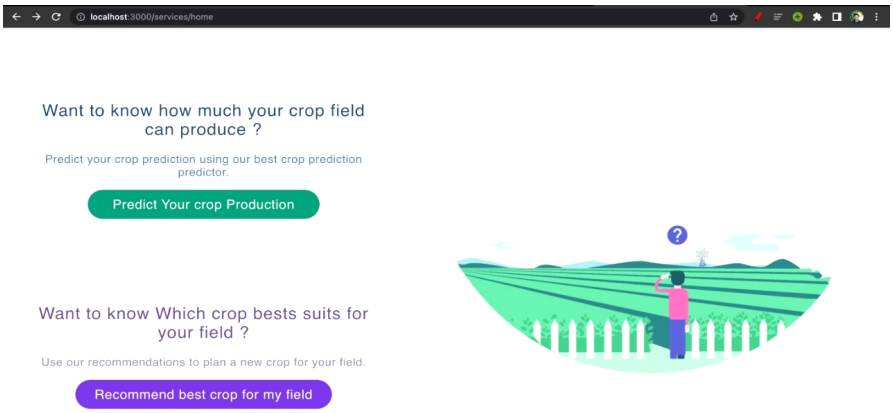
Finally, On implementing the above three algorithms and cross validating the prediction results we come to the conclusion that our model is most accurate in case of Random Forest Algorithm (R2 score : 0.9473978231931719)

Algorithm	Accuracy
Linear Regression	20.3545 %
Decision Tree	92.1678 %
Random Forest	97.3874 %

Regression results

In terms of accuracy Random Forest Algorithm is far better than Linear Regression and significantly more accurate than the decision trees. Thus it can be used to train our predictive model.

Use Case



The user opens the website, he lands on the UI where he gets the two features available:

- Predict your crop production
- Recommend best crop for my field

Predict your crop production:

This feature takes input of the different fields relevant for the crop prediction model to predict and gives out the list of crops with their yield.

Recommend best crop for my field:

This feature can be called an extension to the " Predict your crop production" feature, as it recommends the crops which have high yield production. The recommended crop is one which is at the highest production side when sorted.

Dataset description

	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0
...	...	...	...	...	...	...	...

The dataset contains the tuples having the following column headers:

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

The dataset available can be described by the parameters in the following table.

	Crop_Year	Area	Production
count	246091.000000	2.460910e+05	2.423610e+05
mean	2005.643018	1.200282e+04	5.825034e+05
std	4.952164	5.052340e+04	1.706581e+07
min	1997.000000	4.000000e-02	0.000000e+00
25%	2002.000000	8.000000e+01	8.800000e+01
50%	2006.000000	5.820000e+02	7.290000e+02
75%	2010.000000	4.392000e+03	7.023000e+03
max	2015.000000	8.580100e+06	1.250800e+09

statistical inference of the dataset

The dataset has undergone preprocessing where missing values are found and the tuple was then dropped to get the cleaned data for the training of the predictive model. Also we added a new column named “Yield” which indicates Production per unit Area.

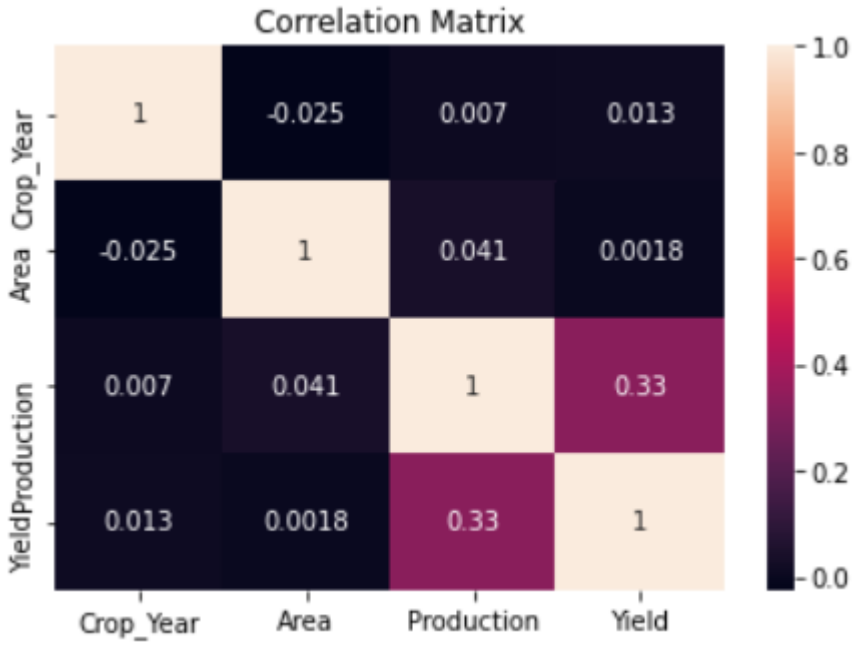
:

	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production	Yield
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0	1.594896
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0	0.500000
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0	3.147059
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0	3.642045
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0	0.229167
5	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Coconut	18168.0	65100000.0	3583.223250
6	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Dry ginger	36.0	100.0	2.777778
7	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sugarcane	1.0	2.0	2.000000
8	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sweet potato	5.0	15.0	3.000000
9	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Tapioca	40.0	169.0	4.225000

After that we dropped an unnecessary column named State_Name. Finally we got the correlation as follows:

	Crop_Year	Area	Production	Yield
Crop_Year	1.000000	-0.025305	0.006989	0.013499
Area	-0.025305	1.000000	0.040587	0.001822
Production	0.006989	0.040587	1.000000	0.330961
Yield	0.013499	0.001822	0.330961	1.000000

Then the heatmap for the above correlation can be drawn as follows:



Correlation Matrix for the dataset

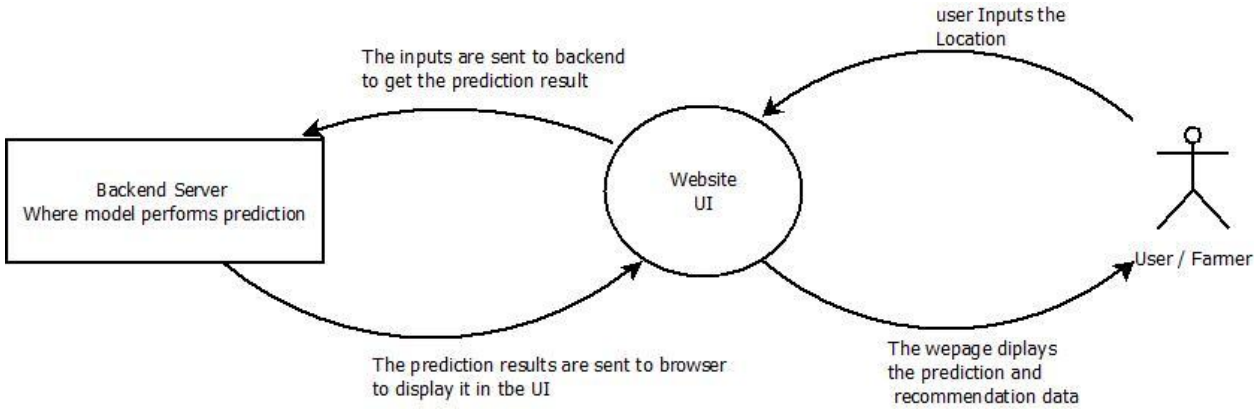
Finally, we splitted our dataset into training dataset and test dataset. where 25% of the available preprocessed dataset was made the test dataset and remaining 75% was made the training dataset.

Thus the dataset was ready to use for training the predictive model.

Proposed System

The crop yield prediction system is going to be a very handy tool for agriculture-related users. The system will let the user input the location (i.e. city, state, etc) and on the basis of it the modal will predict the yield for that location. Also the system is made to recommend the most productive crops for the location.

Software Architecture



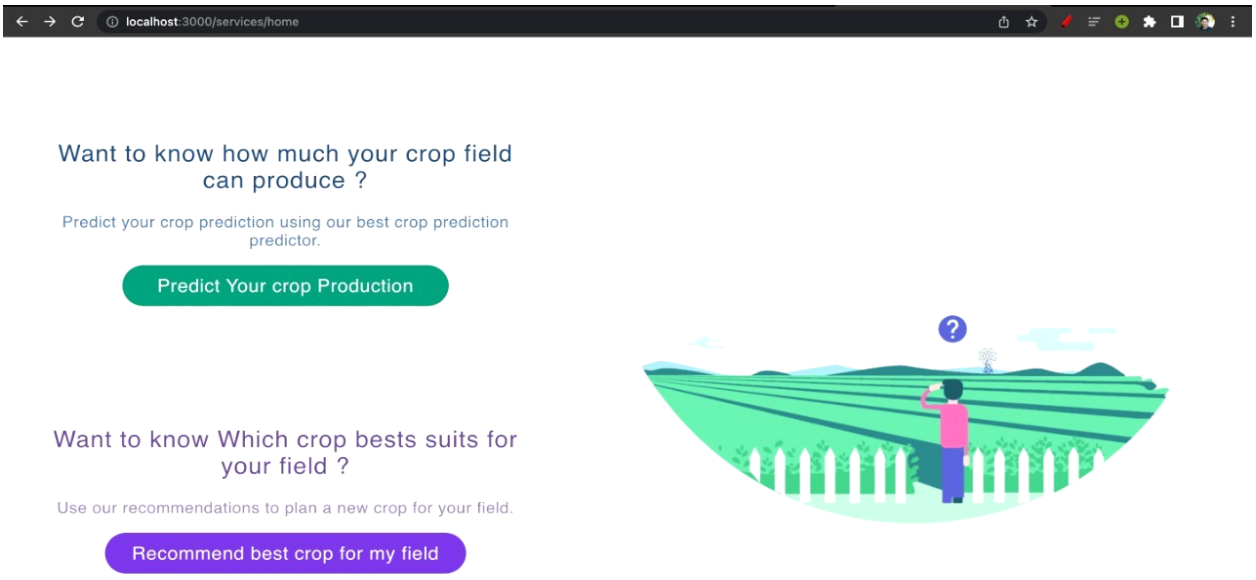
Our Crop yield prediction consists of both frontend and Backend. When the user Inputs the location the data is set to the backend and the prediction model predicts. The prediction is sent to the client / browser where the react framework is used to display the prediction results.

Results

As we have decided to use better and more efficient methods, our accuracy would be far better than the earlier works.

We have come up with a model which is used for crop prediction with its yield using machine learning techniques which were implemented on Jupyter notebook.
The Random Forest Algorithm came to be the most accurate predictor.

A react web-platform is developed for this proposed work which has a user-friendly interface that can be accessed by the user to interact with the model. Once the location-related data is given as input region & season is fetched and compared with the crop dataset, a list of crops suitable for given input condition will be displayed.



The system also has a recommendation system which recommends the crop suitable for the farmer.

Conclusion & Future Scope

This work presents an effective crop prediction system by using past data. Using Machine Learning model and functions datasets are analyzed and the trained model is used to fetch the crop based on region and season. This proposed system helps the farmers to select suitable crops based on season and region of sowing. It will in- turn help the farmers by reducing the loss faced by them and improve net crop yield.

This also contributes to our Indian economy by increasing the yield rate of crop production.

As far as future works are concerned, we can try applying a data independent system i.e. our system should predict with the same accuracy even though data is more or less.
We can refine our prediction Integrating soil details to our advantage, as a parameter.
We can predict whether and how much irrigation is required based on rainfall data.
So on…

References

1. *Crop Yield Prediction based on Indian Agriculture using Machine Learning - 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020*
2. *A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters - Dhivya Elavarasan, · P. M. Durai Raj Vincent, Journal of Ambient Intelligence and Humanized Computing (2021)*
3. *Fuzzy deep learning-based crop yield prediction model for sustainable agronomical frameworks - Dhivya Elavarasan, · P. M. Durai Raj Vincent, Neural Computing and Applications (2021)*
4. *Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey - Nishu Bali, Anshu Singla, Archives of Computational Methods in Engineering (2022)*
5. *Crop Yield Prediction Using Machine Learning Algorithms - Aruvansh Nigam, Saksham Garg , Archit Agrawal, Parul Agrawal, 2019 Fifth International Conference on Image Information Processing (ICIIP)*
6. *A Study on Various Data Mining Techniques for Crop Yield Prediction - Yogesh Gandge, Sandhya, 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*
7. *Crop Prediction using Machine Learning - M.Kalimuthu, P.Vaishnavi, M.Kishore, Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICCSIT 2020)*
8. *An efficient algorithm for predicting crop using historical data and pattern matching technique Anjana, Aishwarya Kedlaya K , Aysha Sana, B Apoorva Bhat, Sharath Kumar, Nagaraj Bhat*
9. *Crop Yield Prediction using Machine Learning Algorithms - Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Prof. Vinu Williams Department of Computer Science and Engineering*
10. *Basso B, Cammarano D, Carfagna E (2013) Review of crop yield forecasting methods and early warning systems. In: The first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics,*
11. *https://www.data.gov.in*
12. *https://en.wikipedia.org/wiki/Agriculture*
13. *Machine Learning Random Forest Algorithm - Javatpoint*