Crop yield prediction using random forest classifier

A minor Project
Report Submitted in partial fulfillment of the
Requirements of VII-Semester for the degree
Of
Bachelor of Technology
In
COMPUTER SCIENCE & ENGINEERING
By

Ashutosh S. Tripathi (17115015)
Under the guidance of
Dr. Naresh Kumar Nagwani
Associate Professor
NIT-Raipur

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY RAIPUR, CG (INDIA)
December, 2020

DECLARATION

I hereby declare that the work described in this thesis, entitled **"Crop yield prediction using random forest classifier"** which is being submitted by us in partial fulfillment for the VIII-Semester of the degree of Bachelor of Technology in the Department of **Computer Science and Engineering** to the National Institute of Technology Raipur is the result of investigations carried out by us under the guidance of **Dr. Naresh Kumar Nagwani (Associate Professor).**

The work is original and has not been submitted for any Degree/Diploma of this or any other Institute/university.

Name of Candidates and Roll No:

Ashutosh S. Tripathi, 17115015

Place: Raipur

Date: 22.04.2020

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

CERTIFICATE

This is to certify that the project entitled "**Crop yield prediction using random forest classifier**" that is being submitted by Ashutosh S. Tripathi (17115015) in partial fulfillment for VIII-semester of the degree of Bachelor of Technology in Computer Science & Engineering to National Institute of Technology Raipur is a record of bonafide work carried out by them under my guidance and supervision.

The matter presented in this project document has not been submitted by them for the award of any other degree elsewhere.

Signature of Supervisor

Dr. Naresh Kumar Nagwani

Associate Professor

Department of Computer Science & Engineering

National Institute of Technology, Raipur (C.G)

H.O.D

Dr. Pradeep Singh Department of C.S.E NIT Raipur (C.G)

Project Coordinator

Dr. Manu Vardhan

Department of Computer Science & Engineering

National Institute of Technology, Raipur (C.G)

ACKNOWLEDGEMENT

We would like to acknowledge my college **National Institute of Technology, Raipur** for providing a holistic environment that nurtures creativity and research-based activities. We express our sincere thanks to **Dr. Naresh Kumar Nagwani,**

Associate Professor CSE Department NIT Raipur, the guide of the project for guiding and correcting throughout the process with attention and care. He has frequently suggested us creative ideas and guided through major hurdles that occurred during the duration of the project.

We would also thank **Dr. Pradeep Singh, Head of the Department** and all the faculty members without whom this project would be a distant reality. We also extend my heartfelt thanks to my family and friends who supported me.

Thank You!!

Ashutosh S. Tripathi (17115015)

Abstract

Over seventy percent of the population in the Indian villages relies on farming, more than eighty percent of whom are either small on marginal in their scope of production. Even then, the livelihood and indeed the food production of the entire country relies on something as fickle as the weather.

Thus, it is paramount for the farmers to know if their crop yield will be reliable or not. The methods used thus far require a good experience in the field and are not completely reliable as the number of factors involved in predicting such a complex variable is too massive for an armchair guess or indeed, even a pen-and-paper computation. In this project, the goal is to delegate this task to the computers, or specifically, to artificial intelligence.

In this project, I use machine learning and the publicly available data like rainfall and maximum temperature in the Kharif season to predict the crop yield. The algorithm used here is the Random forest technique; it is one of the most robust algorithms that involve a decision tree which is also a very popular supervised machine learning algorithm To test its effectiveness, it will be compared to the bagging tree ensemble technique.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

1.1 Overview

In the dawning age of farming, men would find the crops grown in nature and feed the cattle which would lead their small tribe to relative prosperity in comparison to the hunter-gatherer counterparts. In the distant past, the people relied on their own lands to cultivate their food and feed their local community. Even now, India is an agrarian economy and so agricultural practices in India have been absolutely crucial to the country's growth. With the rise of novel innovations, we see an increase in the gross output but an overall decrease in farmer satisfaction.

It would not be an exaggeration to claim that on the whole, technology has been counterproductive to the goal of farmers' happiness and satisfaction. Even modern processes have become inaccessible and abstruse for a common farmer to use. Along with these techniques and the erratic nature of the climate, bountiful crop production has become quite uncertain.So far, deeper meditations on these issues along with the other important factors involved in crop productions have not been entirely fecund. In theory, the possibility of an economic explosion of growth is tangible, yet without a proper direction, the probability seems dubious at best.

Needless to say, there are myriad of possible ways to increase crop productivity, one of the most promising of them is data mining. Essentially, data mining is the process of creating valuable, practical information out of seemingly random, chaotic, and noisy data. In the purest of terms, it turns data into information. Therefore, manual data mining could be quite tedious, this is why experts use dedicated Data Mining software to reorganize, categorize, and summarise their findings of raw, unkempt data. Generally speaking, these softwares allow us to find the underlying pattern between a dozen or so vaguely related variables, each with hundreds or more values.

Thus various links between each variable, patterns, and interlinks can be found with ease. This information is then used to infer the knowledge required to map out the historical repetitions, present trends, and future predictions.

Giving a practical example, should the algorithm project a weak upcoming yield, farmers now have an option to cut their losses. The scope of the application of such software is huge. The sheer number of livelihoods that can be tangibly improved with this is astounding. Before this, the only variable available to predict the yield was the farmer's anecdote about his past yield. The biggest factor in predicting the yield is the weather, though pest problem, plant disease, and the farmer's own hard work are certainly not to be discounted. With accurate software, the farmer will be liable and prepared for creating robust risk avoidance tactics. Therefore, in this project, the idea to predict the yield of the crop is put forward in an attempt to improve the general quality of life for all people but farmers in particular.

1.2 Objectives and Importance of the Project:

The main objectives of the project are listed below:

- To help improve the crop production process.
- To reduce the risks of crop failure.
- To help educate the farming population on the latest technologies.
- To ascertain the precise contribution of each factor involved in predicting the crop production.
- To develop the fields adjacent to farming like weather forecast and crop markets.
- To increase the scope of the machine learning algorithms.
- To help increase the local economic growth.
- To recuperate the loss of economy that has occurred because of the pandemic.
- To compare the Random forest techniques to its peers in terms of accuracy.

1.3 Scope:

As mentioned earlier, this project is largely focused on improving the general methodology of farming and making it more efficient. Thus the brunt of the benefits of this project will be enjoyed by the farmers. With this project, it will be very easy to predict the seasonal output produced by the farmers of any particular crop at any particular season even though the data used here is specifically considering the rice crop in the Kharif season. Thus, the possibility of a monumental increase in output lies just ahead.

However, the possibility of this algorithm being useful in other areas is not to be discounted. With enough data at our disposal, the random forest algorithm is useful at virtually any field that could benefit from its predictive capabilities.

1.4 Motivation

According to a recent survey, over 180 million people in India still suffer from malnutrition. This is an appallingly huge number that can directly be brought down by increasing the overall crop production. Curing hunger in the country is a truly behemoth task that would require years of hard work. I believe this algorithm creates the foundation for this task. By predicting the gross output from each field one can determine the contribution of individual plots in the rice production and hence determine the underlying factors that affect the growth.

In the economic arena, this project can just as easily increase the growth by simply using the variables pertaining to cash crops like cotton. Even a marginal increase in the overall efficiency would mean a significant increase in the overall profits produced because of the high price of such crops.

Thus, the motivations for performing this project are multiple and decidedly important.

1.5: Overview:

In the first chapter, an introduction to the method and its possible applications is provided. The succeeding chapter involves a brief overview of relevant studies. The fourth chapter discussed the methodology whereas the fifth discusses the procedure of creating the program in Rstudio. The subsequent chapters involve the results, future applications and the conclusion drawn from this project respectively.

2. Literature Survey

2.1 Existing work:

Given the grossly useful prospect of this field, significant bit of research has been done in this area. Most of which uses the following algorithms:

- Random forest classifier
- Deep neural network
- The Scikit-Learn Gradient Boosting regressor

P.Priya (2018) created a machine learning algorithm using RStudio to predict the rice crop yield. They used the Random forest technique to predict the yield in the Tamil Nadu region of the country. The dataset used was publicly available records of the government of India in both the Kharif and Ravi seasons from 1997 to 2013.

Andrew Crane-Droesch (2018) describes an approach to yield modeling that uses a semiparametric variant of a deep neural network, which can simultaneously account for complex nonlinear relationships in high-dimensional datasets, as well as the known parametric structure and unobserved cross-sectional heterogeneity. Using data on corn yield from the US Midwest, we show that this approach outperforms both classical statistical methods and fully-nonparametric neural networks in predicting yields of years withheld during model training.

Thomas Van Clompenburg (2020) compiled a list of machine learning and deep learning papers to predict crop yield. 50 machine learning algorithms and 30 deep learning algorithms were compiled to produce a meta-analysis of the topic. Features like rainfall, soil type, and temperature were used to predict the yield. The most widely used algorithm in the ML was neural network while the most widely used algorithm in deep learning was CNN (Convolutional neural network.) The other widely used deep learning algorithms are Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN).

Saeed Khaki (2019) produced a paper that involved Syngenta releasing several large datasets that recorded the genotype and yield performances of 2,267 maize hybrids planted in 2,247 locations between 2008 and 2016 and asked participants to predict the yield performance in 2017 in the

2018 syngenta crop challenge. In this crop challenge, they designed a deep neural network (DNN) approach that took advantage of state-of-the-art modeling and solution techniques. This model was found to have a superior prediction accuracy, with a root-mean-square-error (RMSE) being 12% of the average yield and 50% of the standard deviation for the validation dataset using predicted weather data. With perfect weather data, the RMSE would be reduced to 11% of the average yield and 46% of the standard deviation.

Hajir Almahdi (2020) used several regression models to predict crop yield. He used several models like Gradient Boosting Regressor, Random Forest regressor, SVM, and Decision tree regressor. The Scikit-Learn Gradient Boosting regressor was used as benchmarks and many algorithms were compared to it.

S. Veenadhari (2014) created software called "Crop Advisor" in order to predict the yield of the crop. The software is web-based and is using the C4.5 algorithm. The paper majorly focuses on climate change among other parameters to predict the rate of crop produced at a certain year.

S. No.	Paper Title	Authors	Year	Method used	Limitations
1.	Predicting the yield of crop using a machine learning algorithm	P. Priya, U. Muthaiah, M. Balamurugan	2018	Random forest classifier	The paper doesn't show the results of the findings.
2.	Machine learning methods for crop yield prediction and climate change impact assessment in agriculture	Andrew Crane-Droesch	2018	Deep neural network	The paper was less reliable on warmer climates.
3.	Crop yield prediction using machine learning: A systematic literature review	Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal	2020	N/A. It is a metaanalysis	It is not a singular paper but a metaanalysis
4.	Machine learning approach for forecasting crop yield based on climatic parameters	Veenadhari Suraparaju	2014	C4.5	Very few citations
5.	Crop Yield Prediction Using Deep Neural	Saeed Khaki and Lizhi Wang	2019	DNN	The paper focuses on maize, thus

	Networks				not helpful for rice production
6.	Predicting Crops Yield: Machine Learning Nanodegree Capstone Project	Hajir Almahdi	2020	Gradient Boosting Regressor, Random Forest regressor, SVM, Decision tree regressor, The Scikit-Learn Gradient	No citations

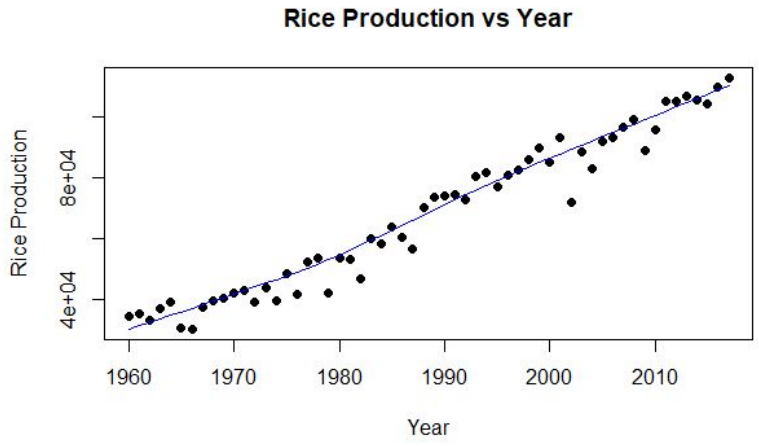
2.2 Summary:

Thus, the aforementioned methods were studied and a few of them were implemented . Their advantages and disadvantages were also considered and evaluated. The properties of various non-decision tree based algorithms were also analysis algorithms were also studied and discussed.

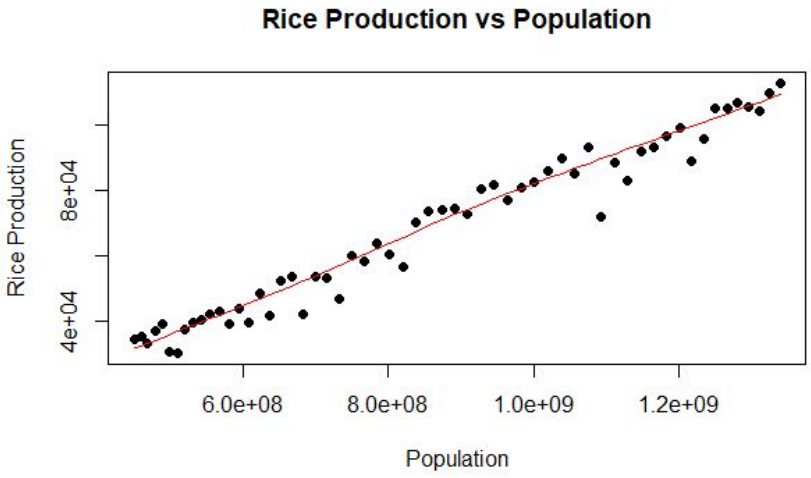
3. Dataset and the code

The dataset includes data from diverse sources but most of it comes from the reports provided by the government of India. This project aims to predict the rice crop yield of the Kharif season in the north Gujarat region of the country. The dataset involves records from the year 1960-2017. The factors considered are annual rainfall, the maximum temperature in the season as it could negatively impact the crop and the previous years' yields.

The first part of building the code was data-preprocessing. The data was cleaned to remove the anomalies. As no major outlier was found, the data was not altered. The second part was altering missing values. Again, no missing values in the data were found thus the data remained unaltered. The data was not noisy thus the methods like regression and clustering were skipped here. The remaining stages of data preprocessing, namely, normalization, discretization, dimensional reduction were considered and found needless for the data. The data was divided into columns responding to each factor discussed above along with the year. In the first stage of the preprocessing, it was found that the rice production increased linearly with time.



As the other variables did not change linearly with time, it was crucial to find the underlying factor involved. Thus, the yearly population data was gathered and it was found that the linear increase can be attributed to the population increase.



Thus, to avoid complications with the data, a “weighted” rice production value was produced with the given formula.

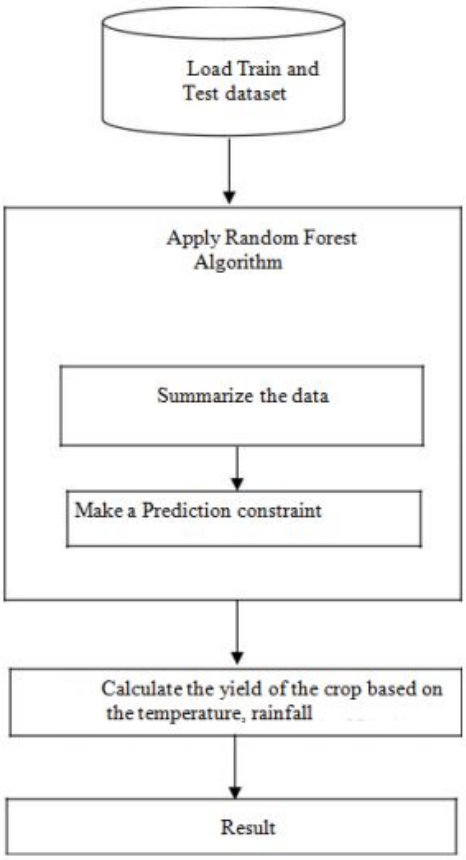
Weighted rice production= (rice production)/(population size)

Thus this value was inputted in the main dataset to produce a more reliable result.

4. Methodology

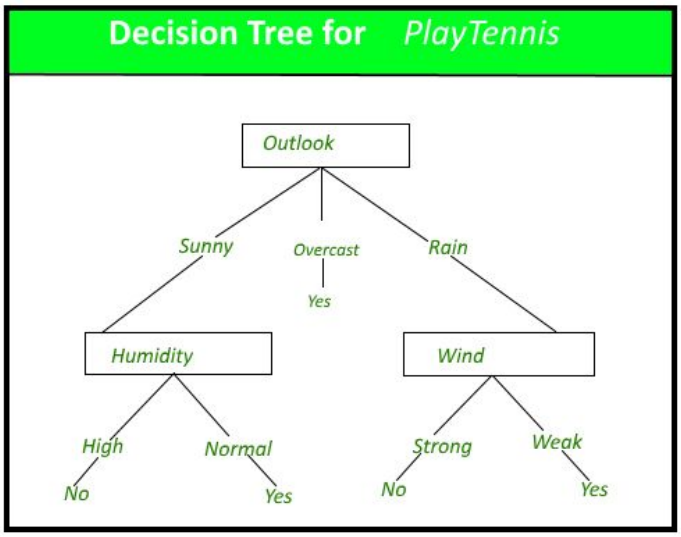
The software on which the computations were performed is Rstudio. It is a leading tool for data science, statistics, and machine learning.

The process of creating the project is shown in this flowchart:



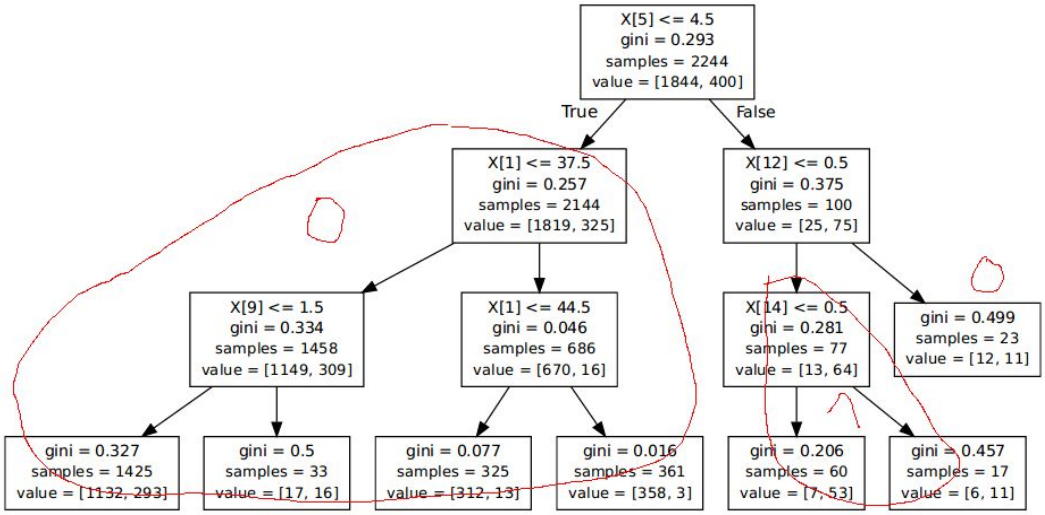
The key concept involved is a regression decision tree. It is a greedy classification approach that involves making a decision at every step to provide the estimated result. The advantage of a decision tree is that it's pictorial representation makes it easier to understand for someone who's not from a computer science or statistics background.

An example of a decision tree is shown below.



A decision tree can be of both regression and classification in nature. The biggest disadvantage a decision tree shows is its high variance. Another noteworthy disadvantage of a decision tree is the time taken to create a tree could be very high as the time complexity increases exponentially with each level. Rstudio libraries have an in-built function to decrease the time-complexity by setting a limit on the tree growth by constricting the tree depth, leaf node volume or a branch node volume.

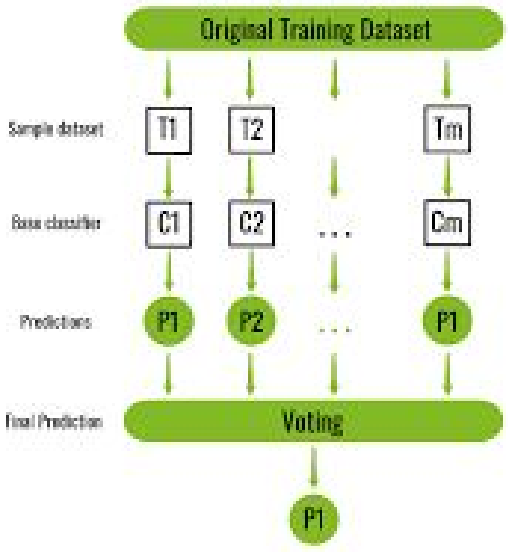
A pruned tree like the one shown below can reduce the time complexity drastically.



Essentially, a pruned tree removes the logically redundant branches to increase the accuracy and decrease the time complexity. Yet despite pruning, the variance of a decision tree still is high compared to its counterparts.

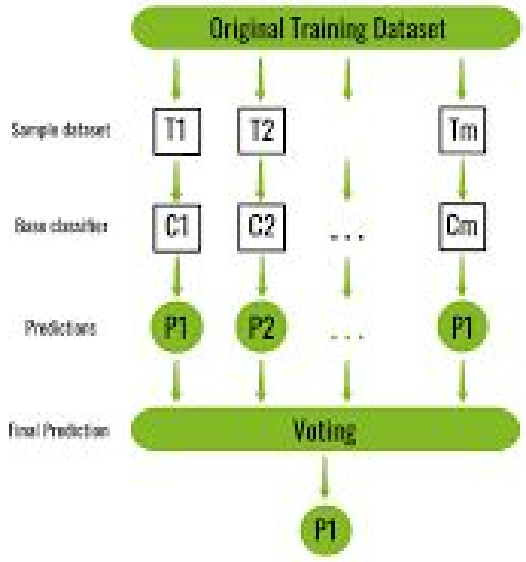
Bagging Classifier

A Bagging classifier is a group meta-assessor that fits base classifiers each on irregular subsets of the first dataset and afterward, it totals their individual forecasts (either by casting a ballot or by averaging) to frame the last expectation. Such a meta-assessor can normally be utilized as an approach to lessen the fluctuation of a discovery assessor (e.g., a choice tree), by bringing randomization into its development methodology and afterward making an outfit out of it.



The above image shows the modus operandi of a bagging tree classifier.

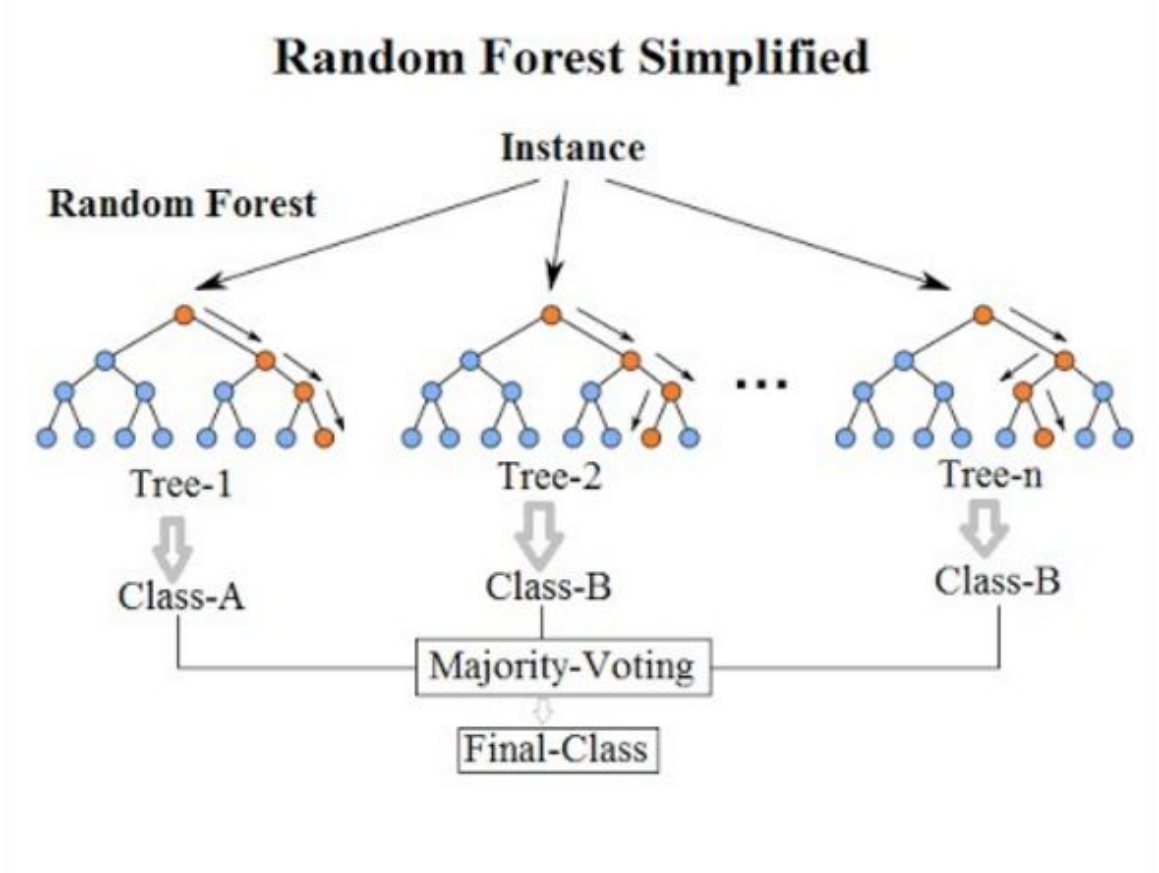
This calculation incorporates a few works from the writing. At the point when arbitrary subsets of the dataset are drawn as irregular subsets of the examples, at that point this calculation is known as Pasting. In the event that examples are drawn with substitution, at that point the strategy is known as Bagging]. At the point when arbitrary subsets of the dataset are drawn as irregular subsets of the highlights, at that point the technique is known as Random Subspaces [3]. At last, when base assessors are based on subsets of the two examples and highlights, at that point the strategy is known as Random Patches.



Random forest classifier:

A random forest is a meta assessor that fits various choice tree classifiers on different sub-examples of the dataset and utilizations averaging to improve the prescient exactness and command over-fitting. The sub-example size is controlled with the max_samples boundary if bootstrap=True (default), in any case the entire dataset is utilized to fabricate each tree.

While a decision tree is easy to understand, the ease is a trade-off as it has a very high variance, thus a few better alternatives are preferred like Random forest classifier and bagging classifier which is an ensemble technique that builds upon the decision tree concept..

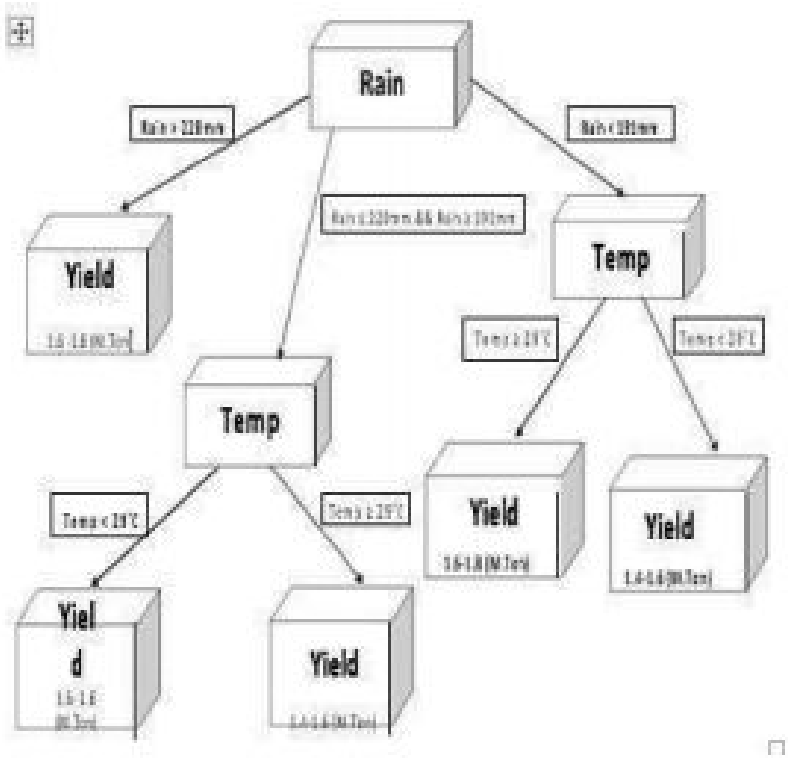


Random forest classifier involves segregating the dataset into several smaller packets of data and making a tree of those. The theory behind this is that if the variance of m trees is n^2 then the variance of all of them combined would the mean n^2/m. Thus having used multiple trees, we will have a so-called random "forest" at our disposal.

4.2 Procedure:

After data preprocessing, the data was randomly split in the ratio of 80:20 using the CA library in Rstudio. Another Library crucial to this project was the randomFor library which containeed the algorithm for both the Random forest classifier and the bagging ensemble technique.

An example of one of the trees that either techniques created is shown below.



As the data was cleft in twain, the first part was the training set and the second part was the test set. The training set will be a randomly chosen subset of the dataset to be used in each of the classifiers to train the algorithms. It's crucial to note that both of the techniques receive the same data to predict the dataset. Once both the bagging ensemble technique and the random forest ensemble technique have been. Having trained the random forest algorithm on the trained on the data, the algorithm was applied to predict the test set. The test set will be the remaining 20 percent of the randomly chosen subset of the data and it will be used to judge the accuracy of each of the algorithms. Please note d that the ensemble techniques have a much better record of showing ,more accurate predictions than the simple decision tree and the pruned decision tree. Hence, those will not be compared to there evidently superior counterparts.

```
test$random<- predict(randomfor, test)
baggingtree<- <- randomForest(wtRiceP~., data = df1, mtry=17)
MSE2random <- mean ((test$random-test$wtRiceP)^2)
test$bag<- predict(baggingtree, test)
MSE2bag <- mean ((test$bag-test$wtRiceP)^2)
```

Having predicted the value, these values were then compared to the actual values by the given formula:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

$\mathbf{RMSD}$ = Root mean square deviation

i = variable i

N = the number of non-missing data points.

x_i = actual observation

$\hat{x}_i$ = estimated observation.

6. Conclusion:

It was found that the MSE of random forest algorithm was roughly 7% more accurate from its bagging counterpart, thus making it more accurate by a reasonable margin

Values	
MSE2bag	0.150607436041486
MSE2random	0.142652682192696

Here, MSE2bag is the mean square error provided by the bagging algorithm and MSE2random is the mean square error provided by the random forest algorithm.

It is hypothesized that we more and more accurate data, the MSE could be reduced even further, yielding more accurate results. With an increase in the data provided, the random forest algorithm is expected to have an even better performance.

This paper shows that farmers can benefit greatly from utilizing machine learning algorithms. With an increase in cooperation between two fields, increased productivity and farmer satisfaction is sure to be found along with higher crop productivity that is set to increase the economic growth and raise the living standards simultaneously.

References:

1. http://www.fao.org/india/fao-in-india/india-at-a-glance/en/#:~:text=70%20percent%20of%20its%20rural,275%20million%20tonnes%20(MT).
2. http://www.ijesrt.com/issues%20pdf%20file/Archive-2018/April-2018/1.pdf
3. https://iopscience.iop.org/article/10.1088/1748-9326/aae159/meta

4. https://research.wur.nl/en/publications/crop-yield-prediction-using-machine-learning-a-systematic-literat
5. https://www.researchgate.net/publication/286582526_Machine_learning_approach_for_forecasting_crop_yield_based_on_climatic_parameters

6. https://www.frontiersin.org/articles/10.3389/fpls.2019.00621/full
7. https://towardsdatascience.com/predicting-crops-yield-machine-learning-nanodegree-capstone-project-e6ec9349f69