

## Machine Learning-Based Models for Crop Yield Prediction: An Analytical Study

A. Oliveira<sup>\*1</sup>, R. Silva<sup>2</sup> & L. Pereira<sup>3</sup>

<sup>\*1</sup>Student, Department of Electrical Engineering, University of São Paulo, São Paulo, Brazil.

<sup>2</sup>Assistant Professor, Department of Mechanical Engineering, University of Porto, Porto, Portugal.

<sup>3</sup>Assistant Professor, Department of Civil Engineering, University of Barcelona, Barcelona, Spain.

---

### ABSTRACT

The agriculture plays a dominant role in the growth of the country's economy. Climate and other environmental changes has become a major threat in the agriculture field. Machine learning (ML) is an essential approach for achieving practical and effective solutions for this problem. Crop Yield Prediction involves predicting yield of the crop from available historical available data like weather parameter, soil parameter and historic crop yield. This paper focus on predicting the yield of the crop based on the existing data by using Random Forest algorithm. Real data of Tami Nadu were used for building the models and the models were tested with samples. The prediction will helps to the farmer to predict the yield of the crop before cultivating onto the agriculture field. To predict the crop yield in future accurately Random Forest, a most powerful and popular supervised machine learning algorithm is used.

**Keywords:** Crop Analysis; Crop Yield; Machine learning; Prediction; Random Forest.

---

### I. INTRODUCTION

Agriculture is the backbone of every economy. In a country like India, which has ever increasing demand of food due to rising population, advances in agriculture sector are required to meet the needs. From ancient period, agriculture is considered as the main and the foremost culture practiced in India. Ancient people cultivate the crops in their own land and so they have been accommodated to their needs. Therefore, the natural crops are cultivated and have been used by many creatures such as human beings, animals and birds. The greenish goods produced in the land which have been taken by the creature leads to a healthy and welfare life. Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are been concentrated on cultivating artificial products that is hybrid products where there leads to an unhealthy life. Nowadays, modern people don't have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food. By analyzing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India there are several ways to increase the economical growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops. Data mining also useful for predicting the crop yield production. Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information.

Data mining software is an analytical tool that allows users to analyze data from many different dimensions or angles, categorize, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The patterns, associations, or relationships among all this data can provide information. Information can be converted into knowledge about historical patterns and future trends. For example, summary information about crop production can help the farmers identify the crop losses and prevent it in future. Crop yield prediction is an important agricultural problem. Each and Every farmer is always tries to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield is primarily depends on weather conditions, pests and planning of harvest operation. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. Therefore, this paper proposes an idea to predict the yield of the crop. The farmer will check the yield of the crop as per the acre, before cultivating onto the field.

### II. LITERATURE SURVEY

A Machine Learning (ML) deals with problems where the relation between input and output variables is not known or hard to obtain. The "learning" term here denotes the automatic acquisition of structural descriptions from examples of what is being described. Unlike traditional statistical methods, ML does not make assumptions about the correct structure of the data model, which describes the data. This characteristic is very useful to

model complex non-linear behaviors, such as a function for crop yield prediction. ML techniques most successfully applied to Crop Yield Prediction (CYP). Supervised Learning algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

#### **Random Forest Classifier:**

Random forest is a most popular and powerful supervised machine learning algorithm capable of performing both classification, regression tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The more trees in a forest the more robust the prediction. Random decision forests correct for decision trees habit of over fitting to their training set

.The data sets considered are rainfall, perception, production, temperature to construct random forest, a collection of decision trees by considering two-third of the records in the datasets. These decision trees are applied on the remaining records for accurate classification. The resultant training sets can be applied on the test data for correct prediction of crop yield based on the input attributes .RF algorithm was used to study the performance of this approach on the dataset. The advantage of random forest algorithm is , Overfitting is less of an issue with Random Forests, unlike decision tree machine learning algorithms. There is no need of pruning the random forest. Random Forest machine learning algorithms can be grown in parallel.

This algorithm runs efficiently on large databases and it has higher classification accuracy .There are three parameters in the random forest algorithm.

- ntree-the name suggests, the number of trees to grow. Larger the tree, it will be more computationally expensive to build models.
- mtry - It refers to how many variables we should select at a node split. The default value is  $p/3$  for regression and  $\sqrt{p}$  for classification and always try to avoid using smaller values of mtry to avoid overfitting.
- nodesize - It refers to how many observations we want in the terminal nodes. This parameter is directly related to tree depth. Higher the number, lower the tree depth. With lower tree depth, the tree might even fail to recognize useful signals from the data.

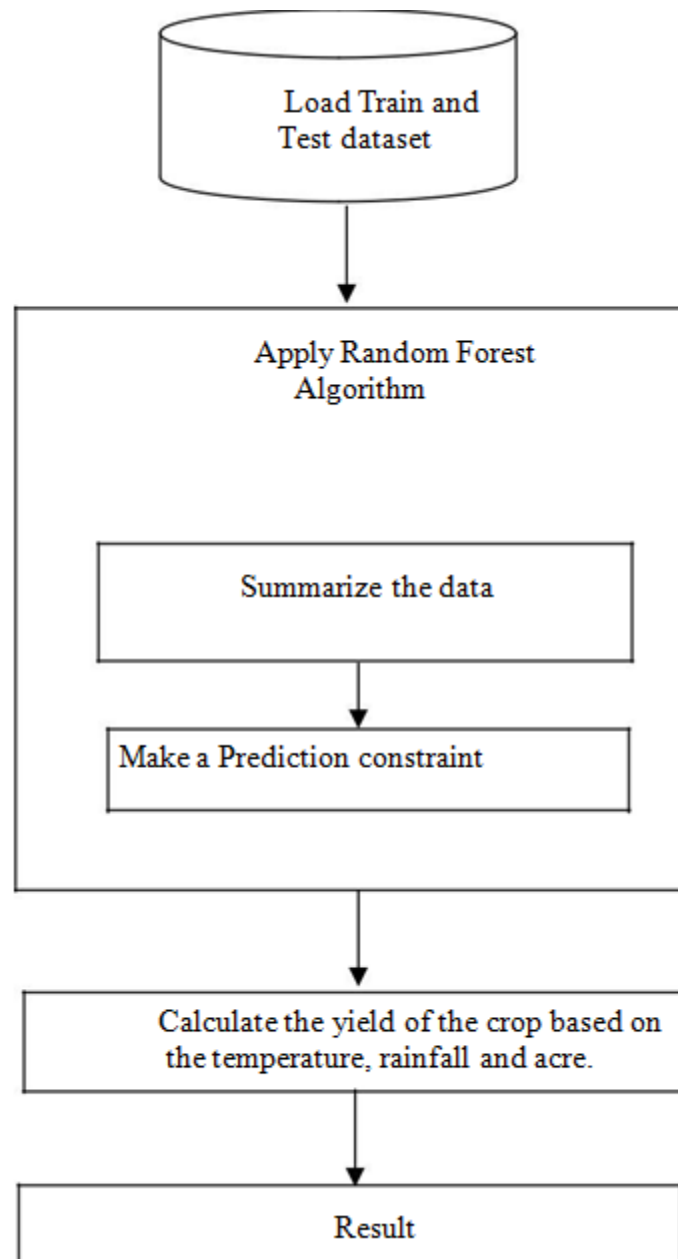
### **III. PROPOSED WORK**

#### **Tool Used**

RStudio was founded by JJ Allaire, creator of the programming language ColdFusion. Hadley Wickham is the Chief Scientist at RStudio. R is the leading tool for statistics, data analysis, and machine learning. It is more than a statistical package; it's a programming language, so you can create your own objects, functions, and packages. It's platform-independent, so it can be used on any operating system and it's free. R programs explicitly document the steps of our analysis and make it easy to reproduce and/or update analysis, which means it can quickly try many ideas and/or correct issues

-

### Work flow



### Dataset Used

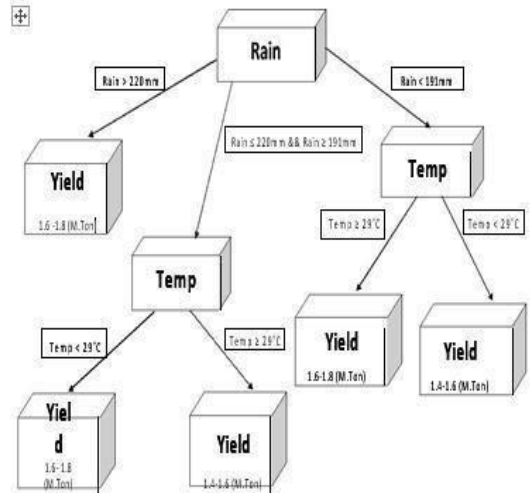
All the datasets used in the research were sourced from the openly accessible records of the Indian Government. This was sourced for the years 1997 to 2013 for different seasons like Kharif and Rabi of rice production. From the vast initial dataset, only a limited number of important factors which have the highest impact on agricultural yield were selected for the present research.

The parameters selected for the present study listed below.

- Rainfall (mm): The total amount of precipitation for
- Kharif and Rabi season of each year of every district.
- Maximum Temperature (degree Celsius): Crop production will definitely have an impact due to maximum temperature for each year of every district was considered for the present research.
- Crop Production (Tonnes): The crop cultivated area in Hectares and production in tonnes for Kharif and Rabi seasons for every year in each selected district of Tamilnadu state was considered for the present research.
- Perception: Perception data for every year in each selected district of Tamilnadu was considered for accurate yield.

### Decision Tree

The Decision tree classifiers uses greedy approach hence an attribute chooses at first step can't be used anymore which can give better classification if used in later steps. Also it overfit the training data which can give poor results for unseen data. So, to overcome this limitation ensemble model is used. In ensemble model results from different models are combined. The result obtained from an ensemble model is usually better than the result from any one of individual models



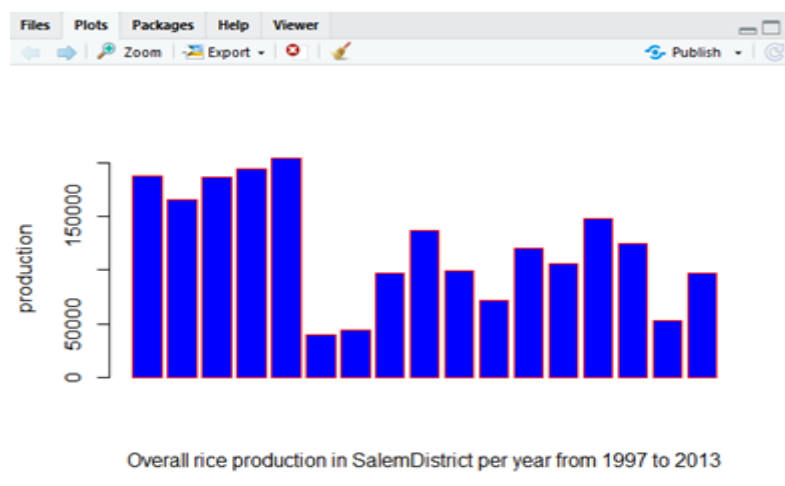
*Example of Decision Tree Implementation*

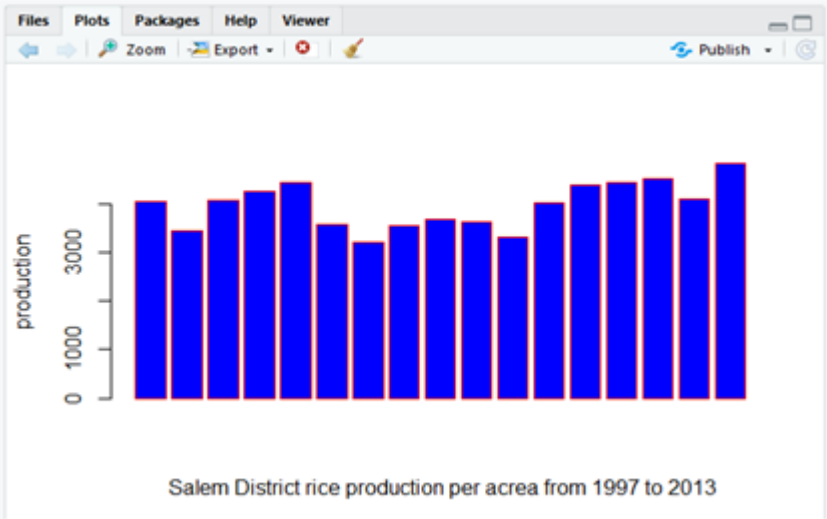
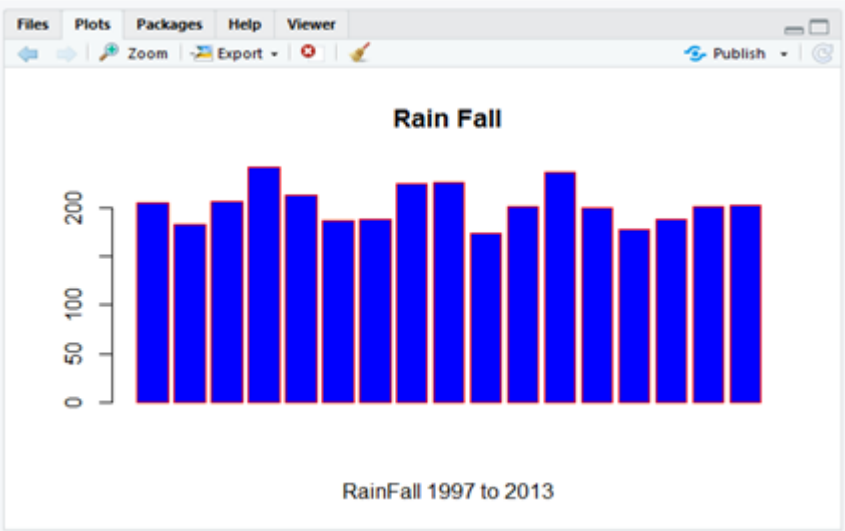
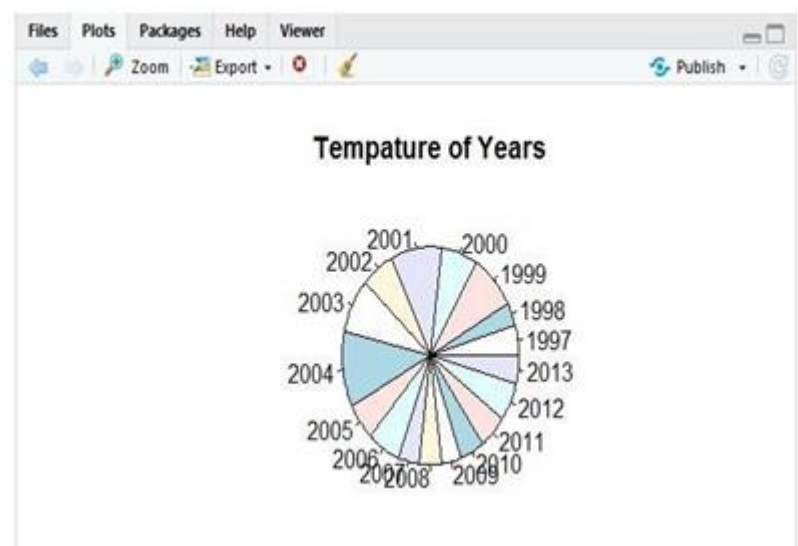
Random Forests is an ensemble classifier which uses many decision tree models to predict the result. A different subset of training data is selected, with replacement to train each tree. A collection of trees is a forest, and the trees are being trained on subsets which are being selected at random, hence random forests. This can be used for classification and regression problems. Class assignment is made by the number of votes from all the trees and for regression the average of the results is used.


### Procedure

According to this algorithm, convert the collected data sets into csv file format and load those data sets. Split the loaded data sets into two sets such as training data and test data in the split ratio of either 67 percentages or 33 percentages that is 0.67 or 0.33. To Separate the training data by class values so that the attribute map to a suitable values and stored in appropriate list. Then calculate Mean and Standard Deviation for needed tuple and then summarize the data sets. Compare the summarized data list and the original data sets calculate the probability. Based on the result the largest probability produced is taken for prediction. The accuracy can be predicted by comparing the resultant class value with the test data set. The accuracy can range from 0% to 100%.

## IV. EXPERIMENTAL RESULTS







```

+ else
+ {
+   ypera <- 2367
+   yield <- a*ypera
+   yield
+ }
+ }
> cat("Rainfall: ",r)
Rainfall: 250
> cat("Temperature: ",t)
Temperature: 36
> cat("Total Acre: ",a)
Total Acre: 25
> cat("Predicted Yield: ",yield)
Predicted yield: 120423.9
>

```

## V. CONCLUSION

The Results shows that we can attain an accurate crop yield prediction using the Random Forest algorithm. Random Forest algorithm achieves a largest number of crop yield models with a lowest models. It is suitable for massive crop yield prediction in agricultural planning. This makes the farmers to take the right decision for right crop such that the agricultural sector will be developed by innovative ideas.

## VI. FUTURE ENHANCEMENT

This paper describes crop yield prediction ability of the algorithm. In future we can determine the efficient algorithm based on their accuracy metrics that will helps to choose an efficient algorithm for crop yield prediction

## VII. REFERENCES

1. Aditya Shastry, H.A Sanjayand E.Bhanushree, "Prediction of crop yield using Regression Technique", International Journal of computing 12 (2):96-102 2017, ISSN:1816-9503
2. E. Manjula , S. Djodiltachoumy, "A Model for Prediction of Crop Yield", International Journal of Computational Intelligence and Informatics, Vol. 6: No. 4, March 2017
3. Mrs.K.R.Sri Preethaa, S.Nishanthini, D.SanthiyaK.Vani Shree , "CropYield Prediction", International Journal On Engineering Technology and Sciences – IJETSTM ISSN(P): 2349-3968, ISSN (O):2349-3976 Volume III, Issue III, March- 2016
4. Askar Choudhury, James Jones, "CROP YIELD PREDICTION USING TIME SERIES MODELS"
5. Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki, "Analysis of agriculture data using datamining techniques: application of big data" Majumdar et al. J Big Data (2017) 4:20 DOI 10.1186/s40537-017-0077-4
6. D. Ramesh and B. Vardhan, "Analysis of crop yield prediction using data mining techniques", International Journal of Research in Engineering and Technology, vol. 4, no. 1, pp. 47-473, 2015.
7. Yethiraj N G , "Applying data mining techniques in the field of Agriculture and allied sciences", Vol 01, Issue 02, December 2012