Crop Prediction by Soil Analysis using Enhanced Random Forest Algorithm for Vidarbha Regions

Submitted in partial fulfilment of the requirements

of the degree of

MASTER OF ENGINEERING

In

COMPUTER ENGINEERING

(Academic Year: 2019-20)

by

Tiwari Priyankar Ravindra Chandrakala

University Registration Number: Thakur /337

Under the Guidance of

Dr. Anand Khandare

Assistant Professor (CMPN Dept.)

University of Mumbai

Department of Computer Engineering (P. G.)

THAKUR TRUSTS
Laxdu Singh Charitable Trust's (Regd.)
THAKUR COLLEGE OF ENGINEERING & TECHNOLOGY
(Autonomous College Affiliated to University of Mumbai, Approved by AICTE & Govt. of Maharashtra)
Institute Accredited by National Assessment and Accreditation Council (NAAC), Bangalore#
Programmes Accredited by National Board of Accreditation (NBA), New Delhi*
Conferred Autonomous Status by University Grants Commission (UGC) for 10 years w.e.f. AY 2019-20.
Among Top 200 Colleges in the Country where Ranked 193rd in NIRF India Ranking 2019 in Engineering College category
*Permanent Affiliated UG Programmes : • Computer Engineering • Electronics & Telecommunication Engineering • Information Technology (w.e.f. A.Y. 2015-16) • Electronics Engineering (w.e.f. A.Y. 2017-18)
*3rd cycle NBA Accredited UG Programmes : • Computer Engineering • Electronics & Telecommunication Engineering • Information Technology (3 years w.e.f. 01-07-2019)
1st cycle of NAAC Accreditation : • "A" Grade for 5 years (w.e.f. 30-10-2017)
A - Block, Thakur Educational Campus, Shyamnarayan Thakur Marg, Thakur Village, Kandivali (East), Mumbai - 400 101.
Tel.: 6730 8000 / 8106 / 8107
Fax : 2846 1890
Email : tcet@thakureducation.org
Website : www.tcetmumbai.in • www.thakureducation.org
tcet
ISO 9001:2015 Certified

Dissertation Approval for M.E.

This is to certify that the dissertation work entitled **" Crop Prediction by Soil Analysis using Enhanced Random Forest Algorithm for Vidarbha Regions "** for M.E. in **Computer Engineering** submitted to University of Mumbai by **"Tiwari Priyankar Ravindra Chandrakala"**, a bonafide student of Thakur college of Engineering and Technology, Kandivali, Mumbai is approved for the award of Master of Engineering Degree in Computer Engineering.

Guide:

Signature: ----------

Name :Dr. Anand Khandare

Assistant Prof, Department of Computer Engineering (P.G.)

Head of the Department :

Signature: ----------

Name :Dr. R. R. Sedamkar

Professor and Head, Department of Computer Engineering (P.G.)

Signature: ----------

Name :Dr. B. K. Mishra

Principal,
Thakur College of Engineering and Technology.

Examiners

1.Signature :----------

Name :

2.Signature :----------

Name :

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

Name: Tiwari Priyankar Ravindra Chandrakala

University Registration Number: Thakur/337

Date:

Place:

Abstract

In the years since its independence, India has made immense progress towards food security. India ranked in the world's five largest producers of over 80% of agricultural produce items. India exported $87 billion worth of agricultural products in 2018, making it the seventh largest agricultural exporter worldwide, and the sixth largest net exporter. India is the fastest growing exporter of agricultural products over a decade period. So, yield prediction is very popular among farmers. Earlier yield prediction was performed by considering the farmer's experience on a particular field, weather and crop. Since farmers don't have knowledge about the presence of the nutrients, Selection of the crop for planting is one of the major challenges faced by farmers. Crop selection is influenced by many factors like the weather, nature of soil, market, etc. Weather and soil type are the major factors which affect the crop yield. Crop yield can be accurately predicted by considering the parameters like nature of the soil, amount of rain, crop characteristics, etc.

There are various methods which can be used to predict crop yield. Artificial Neural Network (ANN) and Machine Learning (ML) are two well-known prediction technique. In this work, prediction of crop yield is done by considering parameters like amount of rainfall, nutrient in soil and amount of soil. In this work Random Forest (RF) is enhanced in terms of accuracy, R^2 Score, Error Rate and Mean Squared Error. Random Forest Algorithm makes decision tree randomly and based on that it predicts the result. For classification it takes majority voting from all decision tree generated and for regression it uses average of all result from generated decision tree. This work presents a system, which uses Random Forest with enhanced performance in order to predict the category of the analysed soil datasets.

In this study, to enhance random forest various parameters have been taken into care for enhancement like reducing error rate, mean squared error, OOB error rate, increasing accuracy, ROC value and AUC. Traditional Random Forest needs number of trees to be entered as input where as enhanced algorithm first checks for number of trees giving the best accuracy and uses that for input which enhances accuracy. For this work, datasets are collected from Department of Agriculture, Government of Maharashtra, which needed a lot of refinement due to missing values. Proposed algorithm has been tested on 10 datasets from different regions to cross verify the accuracy, OOB Error Rate, ROC Value, R^2 Score and compared with ID3 and Traditional Random Forest.

List of Contents

Abstract i
List of Contents ii
List of Figures iv
List of Tables vi
Abbreviations and Symbols vii

Chapter 1. Introduction 1

1.1 Problem Definition 2
1.2 Objective 3
1.3 Motivation 3
1.4 Organization of Dissertation Report 3

Chapter 2. Review of Literature 5

2.1 Introduction 6
2.2 Classification Technique 6
2.3 Regression Analysis 7
2.3.1 Simple Linear Regression 7
2.3.2 Multiple Linear Regression 7
2.4 Background of Ensemble Classification 8
2.5 Random Forest 10
2.5.1 Random Forest Algorithm 11
2.5.2 Outline of Random Forest Algorithm 12
2.5.3 Advantage of Random Forest Algorithm 13
2.5.4 Drawback of Random Forest Algorithm 13
2.5.5 Implementation 13
2.6 Earlier Developments to Random Forest 16
2.7 Random Forest Related work 18
2.8 Random Forest Extension 19
2.9 Application Random Forest 21

Chapter 3. Design Methodology 23

3.1 Introduction 24
3.2 Problem Statement 24

3.3 Architectural Design 24
3.4 Theorotical Framework for Enhanced Random Forest Algorithm 25
3.5 Workflow for Enhanced Random Forest Algorithm 25
3.5.1 Random Forest Classifier and Regression 25
3.5.2 How does this algorithm work 26
3.5.3 Enhanced Random Forests Algorithm 27

Chapter 4. Result and Analysis 29

4.1 Result and Analysis 30
4.1.1 Dataset Description 30
4.2 Result Parameters 31
4.2.1 ROC Curve 32
4.2.2 AUC: Area Under ROC Curve 33
4.2.3 Accuracy 34
4.2.4 Confusion Matrix 35
4.2.5 Area Under Curve 36
4.2.6 Mean Squared Error 37
4.2.7 Precision 37
4.2.8 Racall 38
4.3 Result Obtained by Standard Random forest algorithm 38
4.4 Result Obtained by ID3 Algorithm 42
4.5 Result Obtained by Enhanced Random Forest Algorithm 45
4.6 Comparison 49

Chapter 5. Conclusion 52

Literature Cited 54

Publication 58

Acknowledgement 69

List of Figures

Figure 1: Flowchart of Standard Random Forest Algorithm 14

Figure 2: Performance Measurement of Standard RF 14

Figure 3: Crop Predicted by Standard RF 15

Figure 4: Flow of designed system 24

Figure 5: Random Forest as classification and as regression 25

Figure 6: Flowchart of Enhanced Random Forest 28

Figure 7: Proper Dataset Snap 30

Figure 8: Rainfall Dataset Snap 30

Figure 9: Crop Database Snap 31

Figure 10: Soil sample dataset overview 31

Figure 11: Input for the soil nutrient 32

Figure 12: ROC Curve for the Training Dataset 33

Figure 13: ROC Curve for the Testing Dataset 33

Figure 14: Algorithm Performance 34

Figure 15: ROC Curve 37

Figure 16: Predicted Crop 38

Figure 17: Soil sample dataset overview 38

Figure 18: ROC Curve for the Training Dataset 39

Figure 19: ROC Curve for the Testing Dataset 39

Figure 20: Algorithm Performance for Traditional Random Forest Algorithm 40

Figure 21: Predicted Crop by Standard RF 41

Figure 22: Soil Dataset Overview 42

Figure 23: ROC Curve for the Training Dataset for ID3 42

Figure 24: ROC Curve for the Training Dataset for ID3 43

Figure 25: Performance Measurement for ID3 44

Figure 26: Predicted crop using ID3..... 45

Figure 27: Soil Dataset Overview 45

Figure 28: ROC Curve for the Training Dataset for Enhanced RF 46

Figure 29: ROC Curve for the Training Dataset for Enhanced RF 46

Figure 30 Performance Measurement for Enhanced RF 47

Figure 31: Predicted crop using Enhanced RF 48

Figure 32: Accuracy Comparison..... 50

Figure 33: Error Rate Comparison 50

Figure 34: R^2 Comparison 51

Figure 35: Precision Comparison 51

List of Tables

Table 1: Standard Random Forest Performance 41

Table 2: Performance table for ID3 44

Table 3: Performance table for Enhanced RF 48

Table 4: Comparison of RF and ID3 against Enhanced RF 49

Abbreviations and Symbols

ID3	Iterative Dichotomiser 3
DT	Decision Tree
RF	Random Forest

CHAPTER 1: INTRODUCTION

Introduction

Agriculture is the backbone of the Indian. The agriculture data increases day by day. Since a large population lives in rural areas and is directly or indirectly dependent on agriculture for a living. Outlay from farming forms the main source for the farming community. The essential requirements for harvesting are water resources and ability to buy seeds, fertilizers, pesticides, labour etc. Most farmers raise the required capital by compromising on other essential expenditures, and when it is still insufficient, they resort to credit from sources like banks and private commercial institutions. In such a situation, the repayment is dependent on the success of the harvest. If the harvest fails even once due to several factors, like bad weather pattern; soil type; improper, excessive, and ill-timed application of both fertilizers and pesticides; adulterated seeds and pesticides etc. Most power of soil in nature comes from soil survey efforts. Soil survey, or soil mapping, is the process of determining available nutrients in soil or other holding of the soil cover over a landscape, and mapping them for others to understand and use. Primary data for the soil survey is acquired by area sampling and supported by remote sensing. As the volume of data increase, it requires involuntary way for these data to be extracted when needed. Machine Learning can be used for pretend the next trends of agricultural processes. Every soil is a mixture of these component: Nitrogen, Phosphorus, Potassium, pH Value and Electrical Conductivity. Based on these factors we predict the soil fertility level and crop for a particular soil sample.

In this context, the goal of this thesis is to provide a comprehensive and self-contained analysis of a class of algorithms known as ID3 decision trees and random forests. These methods have proven to be a robust, accurate and successful tool for solving countless of machine learning tasks, including classification, regression, density estimation, manifold learning or semi-supervised learning.

1.1 Problem Definition

A brief study of problems related to maximization of the productivity and prediction of crop yield has been done by going through the related literature review, and with the brief discussions with soil analysts and farmers and broader view of research problem has been gained. Yield prediction is very popular among farmers these days, which particularly contributes to the proper selection of crops for sowing. This makes the

problem of predicting the yielding of crops an interesting challenge. Earlier yield prediction was performed by considering the farmer's experience on a particular field and crop. This work presents a system, which uses Machine Learning techniques in order to predict the category of the analysed soil datasets. The category, thus predicted indicates the yielding of crops.

1.2 Objective

1) To design enhanced algorithm to verify valid soil pattern using semi-automated algorithm.
2) To design enhanced algorithm to characterise soil profile data with soil condition using Random Forest Algorithm.
3) To design and develop an efficient algorithm that would assist in classifying soil with relatively good accuracy.
4) To design and develop an efficient algorithm that would predict crop yield to be plough in soil with relatively good accuracy.
5) Evaluating the performance of proposed algorithm with existing algorithm on different datasets.

1.3 Motivation

Farmers in India, specially Vidarbha region in Maharashtra state faces drought due to which their crop and yielding is getting degraded. They don’t have any idea about availability of nutrient in their field. They use their own experience to plough the crop which have very less success ratio. Due to less success ratio they are unable to pay their loan amount sanctioned for their crop. In unsuccessful for their repayment of the loan amount they attempt to suicide which is a main reason for highly rising ratio in farmers suicide.

To help the farmers to decide the crop to be plough for their benefits I am motivated to build this system. This system collects the data from the soil testing laboratory supported by Department of Agriculture, Government of India. This dataset consists of the available nutrient for farmers’ soil and rainfall for particular region.

1.4 Organization of Dissertation Report

The flow of the project thesis is as follows:

Chapter 1: Introduction consist of introduction, motivation and layout. This chapter gives the introduction of the domain of the project and objective behind the project.

Chapter 2: Literature Survey it is followed by literature survey which consist of different existing systems and the technique used in them. It also covers the limitation of techniques in general.

Chapter 3: Design Methodology design methodology includes the architecture of the system for calculating and classifying the sample using the modified algorithm. It includes flow of the project.

Chapter 4: Result and analyses includes the study of method implemented for predicting crops. Comparative analyses show that current system is better than the system made by combining classification and regression.

Chapter 5: Conclusion concludes the thesis by various facts analyzed from the developed project.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter provides the background knowledge of classification, regression, Random Forest algorithm. To know the current state of art of Random Forest for Classification and Regression in Machine Learning and Artificial Intelligence, this section reviews some of the research works carried out earlier. Random Forest (RF) is an ensemble classifier proposed by Breiman (2001) which consists of many sub-models. The predictions and other quantities of interest are obtained by combining the outputs of all the sub-models.

2.2 Classification Technique

Classification is one of the important decision-making tasks for many real world problems. It is used when an object needs to be classified into a predefined class or group based on attributes of that object (Zhang, 2000)[1]. Generally, there are two types of classification problems: binary problem and multiclass problem. The binary problem is a situation in which an outcome of prediction has to be determined with a decision of Yes or No, whereas a multiple classification problem is a condition in which a predicted result is determined as multiple outcomes (Kraipeerapun, 2009)[3].

Classification organizes data into classes by using predetermined class labels. Classification algorithms normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model, also called a classifier. The model is then applied to predict the class labels for the unclassified objects in the testing data. Typical applications of classification include (but not limited to) credit approval, marketing, and medical diagnosis. Success stories in these areas are too many to enumerate (Khaled *et al.*, 2014)[2].

It has been well recognized that single classifier systems have limited performance (Yan and Goebel, 2004). To boost and improve the performance, ensemble classification which is based on ensemble learning has been used. Ensemble learning uses multiple models to obtain better predictive performance than could be obtained from any of the constituent models (Rokach, 2010; Polikar, 2006; Kuncheva and Whitaker, 2003)[5]. Likewise, ensemble classification employs multiple classifiers and then collectively uses them to identify unlabeled instances.

2.3 Regression Analysis

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However, this can lead to illusions or false relationships, so caution is advisable (Armstrong and Scott, 2012)[4]. In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification (Christopher M. Bishop, 2006)[6].The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems (Waegeman *et al.,* 2008)[7].

2.3.1 Simple Linear Regression

The regression is a statistical technique to determine the linear relationship between two or more variables. Regression is primarily used for prediction and causal inference. In simple regression analysis, one seeks to measure the statistical association between two variables, X and Y. In its simplest (bivariate) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y), as in the formula below:

$$Y= \beta_0 + \beta_1 X + \varepsilon$$

The magnitude and direction of that relation are given by the slope parameter (β_1), and the status of the dependent variable when the independent variable is absent is given by the intercept parameter (β_0). The error term (ε) is the difference between the actual and estimated dependent variable value for any given independent variable values, X_i. The regression coefficient (R_2) shows how well the values fit the data.

2.3.2 Multiple Linear Regression Model

Regression analysis can be expanded to include more than one independent variable. Regressions involving more than one independent variable are referred to as multiple regression. For example, the forecaster might believe that the number of cars sold depends

not only on personal disposable income but also on the level of interest rates. The form of a multiple linear regression model is given by

$$Y_i=\beta_0+\sum_{j=1}^{k}\beta_j X_{ij}+\varepsilon_{i,}$$

where β_0 is the intercept and β_j are known as partial slope coefficients. The partial slope coefficients measure the effect of the independent variable X_i on Y after eliminating the effect of all other independent variables. The intercept captures the expected value of Y when all the independent variables are at level zero.

2.4 Background of Ensemble Classification:

Ensemble classification is an application of ensemble learning to boost the accuracy of classification. Ensemble learning is a machine learning paradigm where multiple models are used to solve the same problem (Rokach, 2010; Polikar, 2006; Kuncheva and Whitaker, 2003). In ensemble classification, multiple classifiers are used and are more accurate than the individual classifiers in the ensemble. A voting scheme is then used to determine the class label for unlabeled instances. A simple and yet effective voting scheme is majority voting (Lam and Suen, 1997). In majority voting, each classifier in the ensemble is asked to predict the class label of the instance being considered. Once all the classifiers have been queried, the class that receives the greatest number of votes is returned as the final decision of the ensemble.

The veto voting is an alternative voting scheme where one single classifier vetoes the decision of other classifiers (Shahzad and Lavesson, 2012; Sun and Dance, 2012). A recent voting scheme by Shahzad and Lavesson (2013) is called trust-based veto voting and is an extension of veto voting. This voting scheme considers the trust of each classifier to determine whether a classifier or set of classifiers can veto the decision.

To achieve optimal results, the classifiers in the ensemble should both be accurate and diverse. An accurate classifier is one that has an error rate better than random guessing. Two classifiers are diverse if they make different errors on new data points. The more diverse the classifiers are, the better the results are. In fact, it has been proven empirically that ensembles tend to yield better results when there is a significant diversity among the models (Kuncheva and Whitaker, 2003). This explains why many ensemble methods seek to promote diversity among the models they combine (Adeva *et al.,* 2005; Brown *et al.,* 2005).

Three widely used ensemble approaches could be identified, namely, boosting, bagging, and stacking. Boosting is an incremental process of building a sequence of classifiers, where each classifier works on the incorrectly classified instances of the previous one in the sequence. AdaBoost (Freund and Schapire, 1997) is the representative of this class of techniques. However, AdaBoost is prone to overfitting. The other class of ensemble approaches is the Bootstrap Aggregating (Bagging) (Breiman, 1996a). Bagging involves building each classifier in the ensemble using a randomly drawn sample of the data, having each classifier giving an equal vote when labeling unlabeled instances. Bagging is known to be more robust than boosting against model overfitting. RF is the main representative of bagging (Breiman, 2001).

Stacking (sometimes called stacked generalization) extends the cross-validation technique that partitions the data set into a held-in data set and a held-out data set; training the models on the held-in data; and then choosing whichever of those trained models performs best on the held-out data. Instead of choosing among the models, stacking combines them, thereby typically getting performance better than any single one of the trained models (Wolpert, 1992). Stacking has been successfully used on both supervised learning tasks (regression) (Breiman, 1996b) and unsupervised learning (density estimation) (Smyth and Wolpert, 1999). Dietterich (2000) conducted an experimental study to compare the performance of three methods for constructing ensembles of classifiers using C4.5 (Quinlan, 1993), namely, bagging (Breiman, 1996a), boosting (Freund and Schapire, 1997), and randomization. With little or no noise in the data, experiments showed that boosting has proved superior to bagging and randomization.

Bagging and randomization demonstrated similar performance, however, with low noise in the data, randomization performed slightly better. Boosting performance seemed to deteriorate with noise. Bagging performance, on the other hand, seemed to improve, for it was able to utilize the noise to produce more diverse classifiers. This finding is consistent with Kuncheva and Whitaker (2003) and Adeva *et al.* (2005) that advocate for diversity to achieve better results. Other experiments showed that, unlike bagging, by increasing the noise rate, randomization failed to produce diverse classifiers. In the next section, we will look at developments that preceded RF.

2.5 Random Forest

RF is an ensemble learning method used for classification and regression. Developed by Breiman (2001), the method combines Breiman's bagging sampling approach (1996a), and the random selection of features, introduced independently by Ho (1995; 1998)[8] and Amit and Geman (1997), in order to construct a collection of decision trees with controlled variation. Using bagging, each decision tree in the ensemble is constructed using a sample with replacement from the training data. Statistically, the sample is likely to have about 64% of instances appearing at least once in the sample. Instances in the sample are referred to as in-bag instances, and the remaining instances (about 36%) are referred to as out-of-bag instances.

Each tree in the ensemble acts as a base classifier to determine the class label of an unlabelled instance. This is done via majority voting where each classifier casts one vote for its predicted class label, and then the class label with the most votes is used to classify the instance.

Breiman (2001) introduced additional randomness during the construction of decision trees using the classification and regression trees (CART) technique. Using this technique, the subset of features selected in each interior node is evaluated with the Gini index heuristics. The feature with the highest Gini index is chosen as the split feature in that node. Gini index has been introduced by Breiman *et al.* (1984)[9].

However, it has been first introduced by the Italian statistician Corrado Gini in 1912. The index is a function that is used to measure the impurity of data, that is, how uncertain we are if an event will occur. In classification, this event would be the determination of the class label (Bader-El-Den and Gaber, 2012). In its general form, it can be calculated as

$$Gini(t)=1-\sum_{i=1}^{N}P(C_i|t)^2$$

where t is a condition, N the number of classes in the data set, and C_i is the i^{th} class label in the data set. In the original paper on RF (Breiman, 2001), it was shown that the RF error rate depends on *correlation* and *strength*. Increasing the correlation between any two trees in the RF increases the forest error rate. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the RF error rate. Such findings seem to be consistent with a study made by Bernard *et al.* (2010)[11] which showed that the error rate statistically decreases by jointly maximizing the strength and minimizing the correlation.

The key advantages of RF over its AdaBoost counterpart are robustness to noise and overfitting (Boinee *et al.,* 2005; Robnik-Šikonja, 2004; Liaw and Wiener, 2002; Breiman,

2001)[12]. Overfitting generally occurs when a model is constructed in such a way that it fits the data more than is warranted. A model which has been overfit will generally have poor predictive performance, as it does not generalize well. By generalization it means how well the model makes predictions for cases that are not in the training set. Hawkins (2004) pointed out that overfitting adds complexity to a model without any gain in performance or, even worse, leads to poorer performance. A classifier that suffers from overfitting is likely to have a low error rate for the training instances (in-bag instances), and a higher error rate for the out-of-bag instances.

2.5.1 Random Forest Algorithm

Random Forest (RF) is an ensemble classifier proposed by Breiman (2001) which consists of many sub-models. The predictions and other quantities of interest are obtained by combining the outputs of all the sub-models. The sub-models for Random Forest are classification and regression trees (CART) which is the key for understanding the Random Forest.

In the past decade, various methods have been proposed to grow a random forest (Breiman, 2001; Dietterich, 2000; Ho, 1998)[15]. Among these methods, Breiman's method (Breiman, 2001) has gained increasing popularity because it has higher performance against other methods (Banfield et al., 2007)[12].

Let D be a training dataset in an M-dimensional space X, and let Y be the class feature with total number of c distinct classes. The method for building a random forest (Breiman, 2001) follows the process including three steps (Baoxun Xu et al., 2012):

Step 1: Training data sampling: use the bagging method to generate K subsets of training data {D1, D2, ..., DK} by randomly sampling D with replacement;

Step 2: Feature subspace sampling and tree classifier building: for each training dataset Di ($1 \leq i \leq K$), use a decision tree algorithm to grow a tree. At each node, randomly sample a subspace Xi of F features ($F << M$), compute all splits in subspace Xi, and select the best split as the splitting feature to generate a child node. Repeat this process until the stopping criteria is met, and a tree hi(Di, Xi) built by training data Di under subspace Xi is thus obtained;

Step 3: Decision aggregation: ensemble the K trees {h1(D1, X1), h2(D2, X2),...hK(DK, XK)} to form a random forest and use the majority vote of these trees to make an ensemble classification decision. (i.e., majority votes for classification, average for regression).

The algorithm has two key parameters, i.e., the number of K trees to form a random forest and the number of F randomly sampled features for building a decision tree. According to Breiman (2001), parameter K is set to 100 and parameter F is computed by F=[log2M + 1]. For large and high dimensional data, a large K and F should be used.
The estimation of the error rate can be obtained based on the training data as follows:

1.1 At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls "out-of-bag", or OOB data) using the tree grown with the bootstrap sample.

2.1 Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate.

2.5.2 Outline of the Random Forest Algorithm:

The process starts with the original data as the input matrix. In the first stage the data are divided into training and test sets. On average 63.2% of all samples from the original data set are placed into the training sets and the remaining 36.8% used as the test set. In Stage 2 the training sets are used to build large collection of decorrelated decision trees that later are used in Stage 3 for classification. In Stage 2 each tree is grown using training data set and starts with a root node where small subsets of input variables are selected randomly (usually according to the square root of the number of variables). This process allows optimizing the splits in the trees within the forest. Finally in Stage 3 each tree within the forest is challenged with the test sets: here each of the bootstrap partitions are run through each of the trees in the forest and the votes are counted. In Stage 4 when all the samples from the test set are run through decision trees the numbers of votes are aggregated to get multi-class classification results. Finally, the algorithm generates a probability distribution of random forests.

1. Randomly select **"k"** features from total **"m"** features.
   Where **k << m**
2. Among the **"k"** features, calculate the node **"d"** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until "**l**" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

2.5.3 Advantages of Random Forest:

1. Accuracy is as good as Adaboost and sometimes better.
2. It is faster than bagging or boosting.
3. It gives useful internal estimates of error, strength, correlation and variable importance.
4. It is simple and easily parallelized.

2.5.4 Drawbacks of Random Forest Algorithm

1. Models in Random Forest which has been overfit will have poor predictive performance as it doesn't generalize well. Generalization means how well model makes prediction for the cases that are not in training set.
2. In Random Forest Algorithm we need to choose number of trees.
3. Large number of attributes for prediction and large number of trees makes algorithm slower.
4. For data including categorical variables with different number of levels, random forests are biased in favour of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

2.5.5 Implementation

Input: Soil Sample Value

Output: Cluster

Algorithm:

1. Randomly select **"k"** features from total **"m"** features.

   Where **k << m**
2. Among the **"k"** features, calculate the node **"d"**.
3. Split the node into **daughter nodes** "**l**".
4. Repeat **1 to 3** steps until "**l**" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**

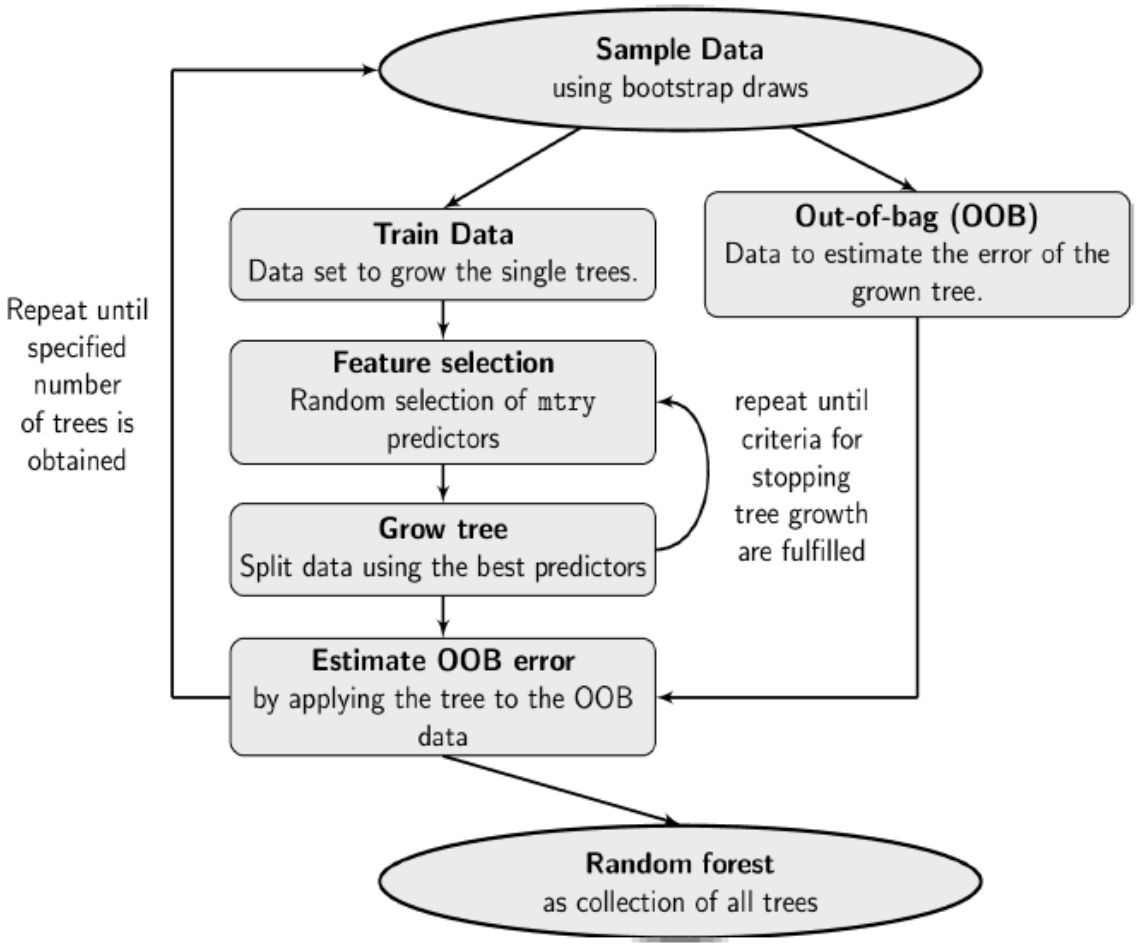


Fig 1: Flowchart of Standard Random Forest Algorithm

```
Select Command Prompt
Microsoft Windows [Version 10.0.17763.503]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\PRIYANKAR R. TIWARI>d:

D:\>cd "ME PROJ"

D:\ME PROJ>cd Final

D:\ME PROJ\Final>py rf_rf.py
Enter value of N : 333
Enter value of P : 7.5
Enter value of K : 507
Enter value of ph : 7.53
Enter value of ec : 0.54
Enter value of Disctrict : Pune
Enter value of year : 2019
Enter value of month : July
Train set
Random Forest:Confusion Matrix:
 [[317  11   0]
 [ 10 339   0]
 [  1  26   0]]
Random Forest OOB error rate : 0.9147727272727273
AUC for random forest: 0.8007215747072657
Accuracy For Random Forest: 0.9318181818181818
Error rate for random forest: 0.06818181818181823
Test set
Random Forest:Confusion Matrix:
 [[80  4  0]
 [ 3 79  0]
 [ 0 10  0]]
Random Forest OOB error rate : 0.9147727272727273
AUC for random forest: 0.7890417867770059
Accuracy For Random Forest: 0.9034090909090909
Error rate for random forest: 0.09659090909090906
Train set
Random Forests Mean squared Error :37863.22479915654
Random Forests r2_score :0.5223538960378842
Random Forest Error Rate :0.47764610396211158
Tesst set
Random Forests Mean Squared error :19369.758205601724
Random Forests r2_score :0.49155448088788034
Random Forest Error Rate :0.5084455191211966
```

Fig 2: Performace Measurement of Standard RF

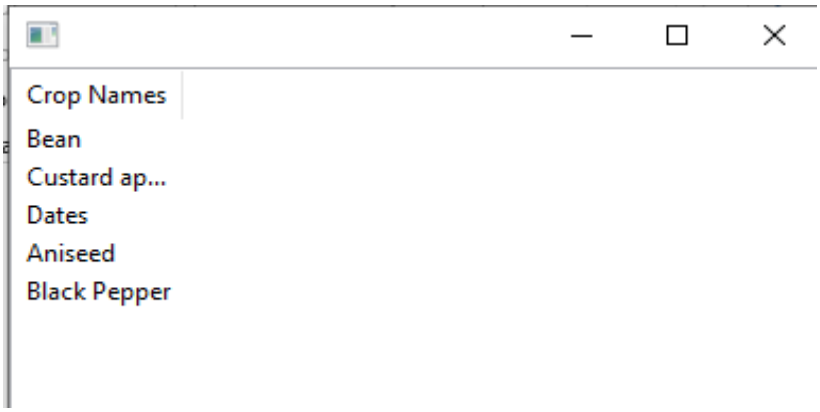


Fig 3: Crop Predicted by Standard RF

The literature for the classification of the soil and prediction of crops are as follows-

In "A Novel Data Mining Approach for Soil Classification"[16] various classification technique are discussed in the paper. The algorithms CART, C4.5 and proposed approach has been studied. Accuracy obtained for C4.5 is more as compared to CART. The proposed system uses gini index and gini ratio. The impurity measure gini index used in CART is biased towards attributes with higher range of values, while information gain used in C4.5 is biased towards attributes with high values. This drawback is removed using the proposed approach. From the paper it is also observed that the manual soil classification being done is very cumbersome and time consuming.

In "Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach"[17][18] Data mining in agriculture field is somewhat a novel research field. Data mining is the process of discovering unknown and likely impressive patterns in large datasets. Steps of data mining such as selection, preprocessing, transformation, data mining and interpretation has been discussed and Naïve bayes and K-nearest neighbor has been used for prediction and analysis. Naive Bayes classifiers can be trained very efficiently in a supervised learning setting and works much better with complex real situations. K-Nearest Neighbor makes predictions based on the outcome of the K neighbors closest to that point. K- nearest Neighbour uses Euclidean distance formula for calculation.

In "Sensible approach for Soil Fertility Management using GIS Cloud"[18] Decision Support System (DSS) using GIS enabled cloud technologies is implemented. The proposed approach of agricultural information development and integration system ensures that complete agricultural related data on cloud database can be integrated into spatial maps through GIS technologies, to organize, accumulate, and administer geospatial data in a cloud database according to individual farmer land information for improving data accuracy of

digital agriculture fertilizer management. IaaS, PaaS and Saas has been discussed with the respective layers used. GIS Cloud server is used to incorporate, integrate, store, update and manage complete information of agriculture.

In “Soil Type Classification and Mapping using Hyperspectral Remote Sensing Data” has demonstrated use of support vector machine algorithm for identification, mapping and classification of various types of soil using high spectral resolution Hyperspectral data. Gaussian Radial Basis Function (RBF) kernel of SVM was used and the accuracy obtained is 71.18.

In “Soil Classification: An Application of Self Organizing Map and k-means”[19] The two unsupervised technique Kohonen Self Organizing Map (SOM) and k-means have been used to classify the soil. Characteristics of soil such as parent material, soil horizontal profile, color of soil, texture of soil and soil depth is listed. This paper predicts the classification of soil and gives information about the plants to be cultivated in specify type of soil.

2.6 Earlier Developments to Random Forest

Ho (1995) proposed a method to overcome a fundamental limitation on the complexity of decision tree classifiers derived with traditional methods. Such classifiers cannot grow to arbitrary complexity without sacrificing the generalization accuracy on unseen data. The proposed method uses oblique decision trees which are convenient for optimizing training set accuracy. The essence of the method is to build multiple trees in randomly selected subspaces of the feature space. The trees generalize their classification in complementary ways, and their combined classification can be monotonically improved.

Amit and Geman (1997) proposed a shape recognition approach based on the joint induction of shape features and tree classifiers. Because of virtually infinite number of features, they reached the conclusion that no classifier based on the full feature set could be evaluated as it was impossible to determine a priori which features were informative. Due to the number and nature of features, standard decision tree construction based on a fixed length feature vector was not feasible. An alternative approach would be to entertain a small random of sample features at each node, constrain their complexity to increase with tree depth, and grow multiple trees. Terminal nodes contain estimates of the corresponding posterior distribution over shape classes. By sending the image down and aggregating the resulting distribution, the image can be classified.

In another paper, Ho (1998) proposed a method to solve the dilemma between overfitting and achieving maximum accuracy. This was done by constructing a decision tree-

based classifier that maintained the highest accuracy on training data and at the same time, improved on generalization accuracy as it grows in complexity. The classifier consisted of multiple trees constructed systematically by pseudo-randomly selecting subsets of components of the feature vector, that is, trees constructed in randomly chosen subspaces. When empirically tested against publicly available data sets, the subspace method proved its superiority when compared to single-tree classifiers and other forest construction methods. The next section introduces RF which is an ensemble method that combines existing techniques in order to construct a collection of decision trees with controlled variation.

Over the past decade, some research was invested in boosting the performance of RF. One of the earliest to be reported is by Latinne *et al.* (2001)[2]. A method based on the McNemar non-parametric test of significance was proposed. The method a priori determines the minimum number of trees in the RF to use in order to obtain prediction accuracy comparable to the one obtained with larger ensembles. In addition to maintaining accuracy with fewer trees, the method significantly improves classification speed and reduces memory costs.

Robnik-Šikonja (2004) investigated new ways to improve the performance of RF[20]. By using several attribute evaluation measures instead of just one, the correlation between trees is decreased without any loss in their strength. Another way to improve the performance of RF is to change the voting method. Instead of using majority voting, weighted voting is used. With this voting technique, internal estimates are used to identify instances most similar to the instance being labeled. The votes of the corresponding trees are then weighted with the strength they demonstrate on these near instances. Improvements were demonstrated on several classification data sets.

Tsymbal *et al.,* (2006)[21] found a way to improve the performance of RF on some data sets by replacing majority voting with more sophisticated dynamic integration techniques. Three techniques were used: Dynamic Selection (DS), Dynamic Voting (DV), and Dynamic Voting with Selection (DVS). Using DV and DVS integration strategies, experimental studies showed that dynamic integration was able to improve the accuracy of RFs on 12 out of 27 data sets.

Amaratunga *et al.,* (2008)[23] investigated the significance decline in RF when the number of features is large and the number of truly informative features is small (as in the DNA microarray data set). The proposed novel and simple approach was to pick the eligible subsets of features to split each node by weighted random sampling instead of simple random

sampling, with the weights tilted in favor of the informative features. The approach demonstrated superior performance when applied to several actual microarray data sets.

Saffari *et al.,* (2009)[21] proposed a novel online RF algorithm to remedy the limitations of the off-line algorithm which has limited usability for many practical problems. Ideas from online bagging, extremely randomized forests, and online decision tree growing procedures were combined to produce the online version of the algorithm. To boost performance, temporal weighting scheme for adaptively discarding some trees based on their out-of-bag error was added. Experiments have shown that the performance of the online algorithm proved comparable to the off-line version.

Bader-El-Den and Gaber developed an approach to enhance the accuracy of RF by using genetic algorithms (Goldberg, 1989)[23]. The approach was called genetic algorithm-based random forest (GARF) (Bader-El-Den and Gaber, 2012)[22]. Experiments have shown that GARF outperformed other state-of-the-art classification techniques including AdaBoost.

2.7 Random Forest Related work

Random forests (Breiman, 2001) were originally conceived as a method of combining several CART (Breiman *et al.,* 1984) style decision trees using bagging (Breiman, 1996a). Their early development was influenced by the random subspace method of Ho (1998), the approach of random split selection from Dietterich (2000) and the work of Amit and Geman (1997) on feature selection. Several of the core ideas used in random forests are also present in the early work of Kwokt and Carter (1988) on ensembles of decision trees.

In the years since their introduction, random forests have grown from a single algorithm to an entire framework of models (Criminisi *et al.,* 2011), and have been applied to great effect in a wide variety of fields (Criminisi and Shotton, 2013; Shotton *et al.,* 2011; Cutler *et al.,* 2007; Prasad *et al.,* 2006; Svetnik *et al.,* 2003). In spite of the extensive use of random forests in practice, the mathematical forces underlying their success are not well understood. The early theoretical work of Breiman (2004) for example, is essentially based on intuition and mathematical heuristics, and was not formalized rigorously until quite recently (Biau, 2012). There are two main properties of theoretical interest associated with random forests. The first is consistency of estimators produced by the algorithm, which roughly guarantees the convergence to an optimal estimator as the dataset grows infinitely large. The second one is rates of convergence; the consistency which surprisingly, has not yet been established even in Breiman's original algorithm.

Theoretical papers typically focus on stylized versions of the algorithms used in practice. An extreme example of this is the work of Genuer (2012; 2010), which studies a model of random forests in one dimension with completely random splitting. In exchange for simplification researchers acquire tractability, and the tact assumption is that theorems proved for simplified models provide insight into the properties of their more sophisticated counterparts, even if the formal connections have not been established.

An important milestone in the theory of random forests is the work of Biau *et al.* (2008), which proves the consistency of several randomized ensemble classifiers. Two models studied in Biau *et al.* (2008) are direct simplifications of the algorithm from Breiman (2001), and two are simple randomized neighbourhood averaging rules, which can be viewed as simplifications of random forests from the perspective of Lin and Jeon (2006).

More recently Biau (2012) has analyzed a variant of random forests originally introduced in Breiman (2004) which is quite similar to the original algorithm. The main differences between the model in Biau (2012) and that of Breiman (2001) are in how candidate split points are selected, and that the former requires a second independent data set to fit the leaf predictors. While the problem of consistency of Breiman's algorithm remains open, some special cases have proved tractable. In particular, Meinshausen (2006) has shown that a model of random forests for quantile regression is consistent and Ishwaran and Kogalur (2010) have shown the consistency of their survival forests model. Denil *et al.* (2013) have shown the consistency of an online version of random forests. The next section explores some extensions to RF. With one common goal in mind, they were mainly developed in order to further boost its performance.

2.8 Random Forest Extensions

Over the past decade, some research was invested in boosting the performance of RF. One of the earliest to be reported is by Latinne *et al.* (2001). A method based on the McNemar non-parametric test of significance was proposed. The method a priori determines the minimum number of trees in the RF to use in order to obtain prediction accuracy comparable to the one obtained with larger ensembles. In addition to maintaining accuracy with fewer trees, the method significantly improves classification speed and reduces memory costs.

Robnik-Šikonja (2004) investigated new ways to improve the performance of RF. By using several attribute evaluation measures instead of just one, the correlation between trees is

decreased without any loss in their strength. Another way to improve the performance of RF is to change the voting method. Instead of using majority voting, weighted voting is used. With this voting technique, internal estimates are used to identify instances most similar to the instance being labeled. The votes of the corresponding trees are then weighted with the strength they demonstrate on these near instances. Improvements were demonstrated on several classification data sets.

Tsymbal *et al.,* (2006) found a way to improve the performance of RF on some data sets by replacing majority voting with more sophisticated dynamic integration techniques. Three techniques were used: Dynamic Selection (DS), Dynamic Voting (DV), and Dynamic Voting with Selection (DVS). Using DV and DVS integration strategies, experimental studies showed that dynamic integration was able to improve the accuracy of RFs on 12 out of 27 data sets.

Amaratunga *et al.,* (2008) investigated the significance decline in RF when the number of features is large and the number of truly informative features is small (as in the DNA microarray data set). The proposed novel and simple approach was to pick the eligible subsets of features to split each node by weighted random sampling instead of simple random sampling, with the weights tilted in favor of the informative features. The approach demonstrated superior performance when applied to several actual microarray data sets.

Saffari *et al.,* (2009) proposed a novel online RF algorithm to remedy the limitations of the off-line algorithm which has limited usability for many practical problems. Ideas from online bagging, extremely randomized forests, and online decision tree growing procedures were combined to produce the online version of the algorithm. To boost performance, temporal weighting scheme for adaptively discarding some trees based on their out-of-bag error was added. Experiments have shown that the performance of the online algorithm proved comparable to the off-line version.

Bader-El-Den and Gaber developed an approach to enhance the accuracy of RF by using genetic algorithms (Goldberg, 1989). The approach was called genetic algorithm-based random forest (GARF) (Bader-El-Den and Gaber, 2012). Experiments have shown that GARF outperformed other state-of-the-art classification techniques including AdaBoost.

Xu, Huang, Williams, and Ye (2012) proposed a hybrid weighted RF algorithm for classifying high dimensional data. It was called hybrid because different decision tree algorithms including C4.5, CART, and CHAID were used to build the trees in the RF. The

hybrid RF was tested on eight high-dimensional data sets. When compared with the traditional RF, results showed that the hybrid approach consistently performed better.

2.9 Application of Random Forest

Over the past decade, many applications of RF were developed in virtually all disciplines, and new applications are yet to be uncovered. The ones chosen in this section are by no means exhaustive as there are many. In this part, some of the interesting applications are discussed.

In Ecology, Cutler *et al.* (2007) compared the accuracies of RF and four other commonly used statistical classifiers using various species data collected from multiple locations in the USA. The results demonstrated RF's superiority over the other techniques.

In Medicine, Klassen *et al.* (2008) conducted some experiments to explore several attribute selection methods with RF that precisely classified cancer using a published benchmark data set. Experimental results showed that RF performed well for microarray data in terms of speed and accuracy with several different gene sets.

Hu (2009) applied RF to study the prediction of pathologic complete response in breast cancer. Results showed that the feature selection scheme of RF was able to identify important genes of biological significance.

In Astronomy, Gao *et al.* (2009) conducted some experiments on multi-wavelength data classification. Results showed that RF proved effective for astronomical object classification. RF has proved to be superior due to its own virtues in classification, feature selection, feature weighting, and detection of outliers.

In Autopsy, Flaxman *et al.* (2011) introduced a new computer-coded verbal autopsy (CCVA) method using RF to predict cause of death. This was done by training RF to distinguish between each pair of causes, and then combining the results through a novel ranking technique. The new method outperformed physician-certified verbal autopsy and was recommended for analyzing past and current verbal autopsies.

In Traffic and Transport Planning, Zaklouta *et al.* (2011) used K-d trees and RFs to classify 43 types of traffic sign using different size histogram of oriented gradients (HOG) descriptors and distance transforms. Results showed that RFs outperformed K-d trees by achieving a classification rate of 97.2% and 81.8% on HOG and distance transforms, respectively.

In Agriculture, Löw *et al.* (2012) used a combination of RF and support vector machine (SVM) classifiers to improve crop classification accuracy and to provide spatial information on map uncertainty. Results showed that the feature selection merit of RF improved the performance of SVM. Using this hybrid classifier improved classification accuracy compared with single classifiers and user's and producer's accuracy.

In Bioinformatics and Computational Biology, Boulesteix *et al.* (2012) amalgamated 10 years of RF development. Practical aspects of RF including selection of parameters, available RF implementations, important pitfalls, and biases of RF and its variable importance measures were covered. This also surveyed recent developments relevant to Bioinformatics as well as some representative examples of RF applications in this domain.

CHAPTER 3. DESIGN METHODOLOGY

3.1 Introduction

Improving accuracy in classification and prediction has been grasping a lot of attention from many researchers all over the world. Random Forest is a new approach to data exploration, data analysis, and predictive modelling. This research work focuses on improving the performance of random forest. On the whole, in various aspects, the findings of the research improve the accuracy of the random forest algorithm for resolving many real world classification and regression problems effectively.

3.2 Problem Statement

Accuracy of classification is one of the important features. To improve accuracy, various strategies have been identified. Ensemble methods are one of the renowned techniques to improve classification accuracy. Ensembles are learning techniques that build a set of classifiers and classify new datasets on the basis of their vote of prediction. Random Forest is the most common ensemble learning algorithm used to improve the classification and prediction accuracy. In this research, accuracy improvement of random forest for classification and regression is carried out in which experiments were performed with different datasets.

3.3 Architectural Design

The designed system works in four steps to process the data and generate the result. The process includes building forest, classifying soil sample, predicting rain fall and finally predicting crop to plough. The figure below shows the architectural design of the designed system.

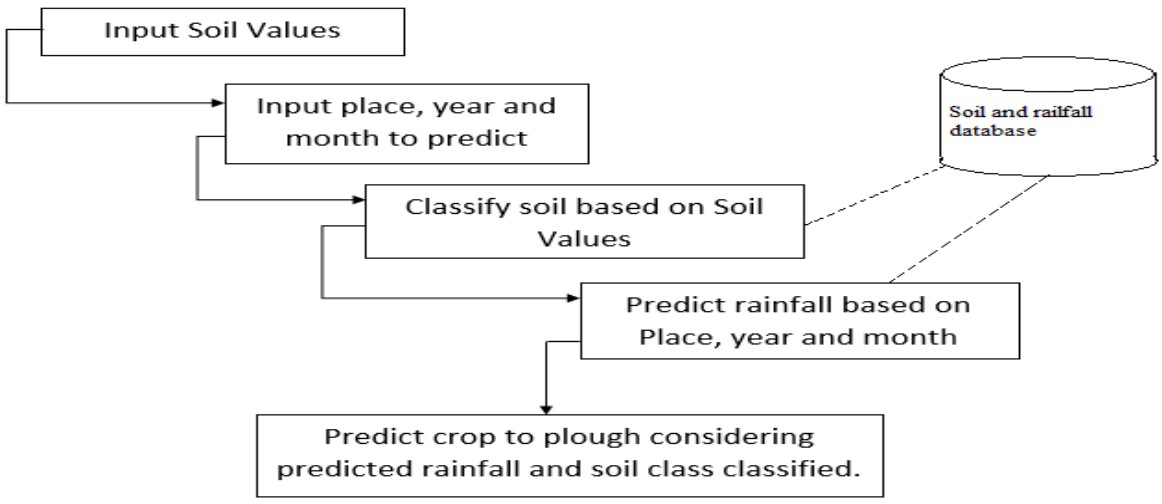


Fig 4: Flow of designed system

3.4 Theoretical Frame Work of Enhanced Random Forest

Random forest algorithm tends to use a simple random sampling in which equal chances are given for all the observations in building their decision trees. This will decrease the classification accuracy of the individual tree in the forest when there are a large proportion of bad observations (noisy, outlier and non-informative) present in the dataset. This research aims to enhance Random Forest algorithm by applying regression method using Random Forest algorithm after the classification of the data to predict the value.

3.5 Workflow for Enhanced Random Forest

Prediction based on Enhanced Random forest algorithm, in this system is used for the predicting crop based on soil sample, nutrient available in soil and rainfall in particular region.

The flow of the system is displayed in blow diagram

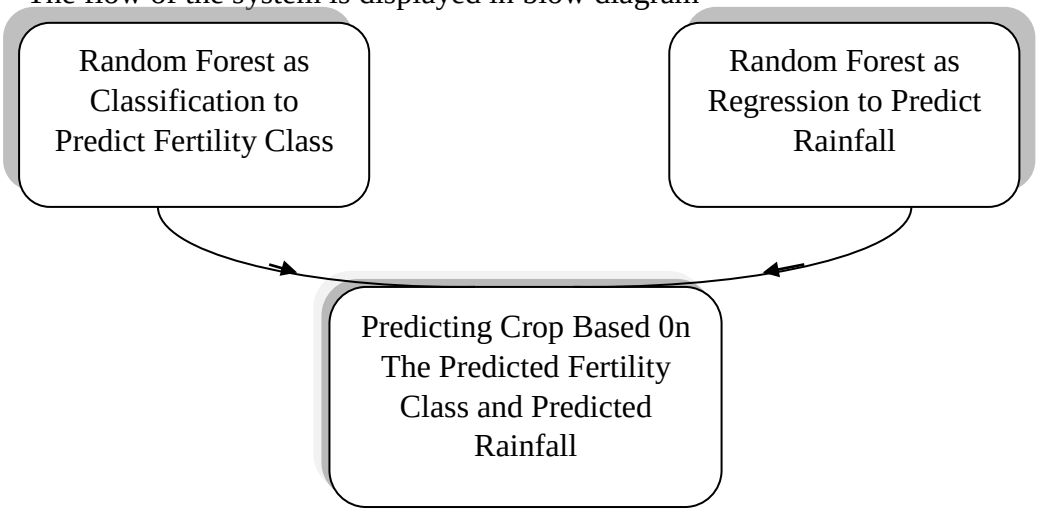


Figure 5: Random Forest as classification and as regression

3.5.1 Random Forest Classifier and Regression

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a

classification problem, each tree votes and the most popular class is chosen as the final result.

The RF regression algorithm is an ensemble-learning algorithm that combines a large set of regression trees. A regression tree represents a set of conditions or restrictions that are hierarchically organized and successively applied from a root to a leaf of the tree [3]. The RF begins with many bootstrap samples that are drawn randomly with replacement from the original training dataset. A regression tree is fitted to each of the bootstrap samples. For each node per tree, a small set of input variables selected from the total set is randomly considered for binary partitioning. The regression tree splitting criterion is based on choosing the input variable with the lowest Gini Index, i.e. $I_G(t_{X(xi)})=1-\sum f(t_{X(xi)},j)^2$, where $f(t_{X(xi)}, j)$ is the proportion of samples with the value x_i belonging to leave j as node t. The predicted value of an observation is calculated by averaging over all the trees[3]. Two parameters need to be optimized in the RF: the number of regression trees (*ntree*; default value is 500 trees) and the number of input variables per node (*mtry*; default value is 1/3 of the total number of variables).

3.5.2 How does this algorithm work?

It works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree.

2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.

Each tree is grown to the largest extent possible. There is no pruning.
The above algorithm generates forest by following above algorithm which is enhanced to improve performance. After building forest, classification and regression is applied as below.

1. Draw n_{tree} bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when, m_{try} = p, the number of predictors.)
3. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

3.5.3 The Enhanced Random Forests Algorithm:

The standard algorithm has two key parameters, *i.e.*, the number of *n* trees to form a random forest and the number of *F* randomly sampled features for building a decision tree. According to Breiman (2001), parameter *K* is set to 100 and parameter *F* is computed by F=[log2M + 1].

To enhance the algorithm, the samples feature should not be limited. For this, in enhanced algorithm number of F max_feature is randomly is entered by algorithm and the best one is selected for the system.

Also the standard algorithm need to be entered the number of "*n*" Trees to a random forest and the larger the number of the trees slower the speed of algorithm. Thus to resolve this enhanced algorithm is implemented with random function which randomly choose the number of trees and checks the result for each and select the number of tree which has the best result and uses that for all the further steps.

1. Randomly select **"k"** features from total **"m"** features.

   Where **k << m**
2. Among the **"k"** features, check for every feature and select **"f"** best feature.

3. For the feature **"f"** calculate the node **"d"**
4. Split the node **"d"** into daughter node **"l"** using best split.
5. Repeat **1 to 3** steps until "**l**" number of nodes has been reached.
6. Randomly put the value of number of trees and select **"n"** the number giving best result
7. Build forest by creating **"n"** number of trees.

The above algorithm generates forest by following above algorithm which is enhanced to improve performance. After building forest, classification and regression is applied as below.

1. Draw n_{tree} bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when, $m_{try} = p$, the number of predictors.)
3. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

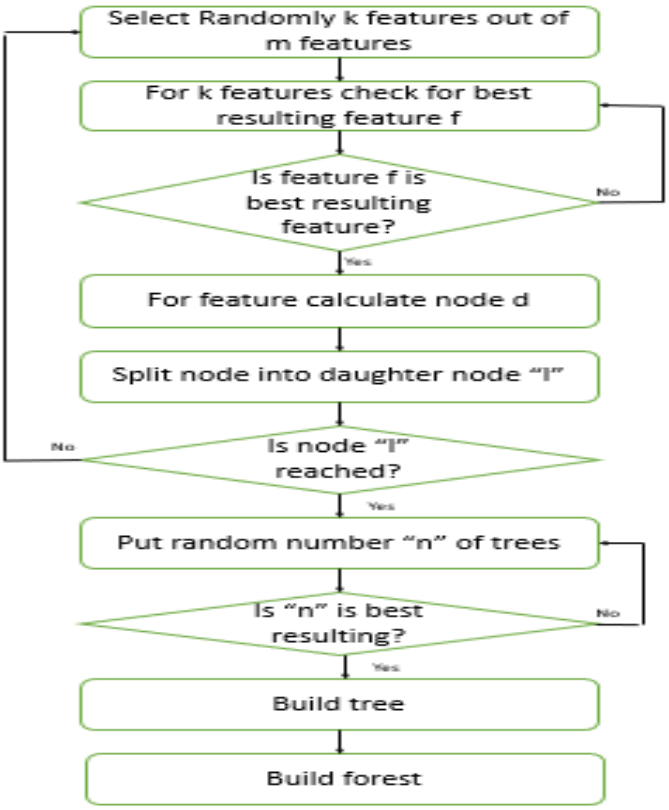


Fig 6: Flowchart of Enhanced Random Forest

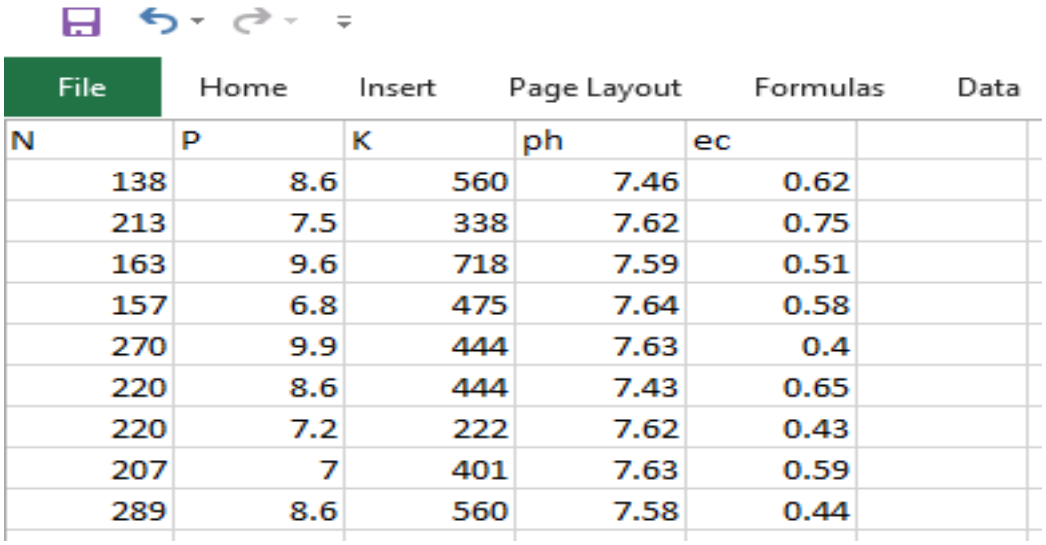
CHAPTER 4. Result and Analysis

4.1 Result and Analysis

This chapter discuss the dataset used and result obtained from the modified algorithm. It also compares it with the existing algorithm. Result obtained from the algorithm have been checked on different types of the datasets.

4.1.1 Dataset Description

Datasets used in this system numeric value which consist of the nutrient value in different unit and rainfall in different area is recorded in mm (Millimetre).

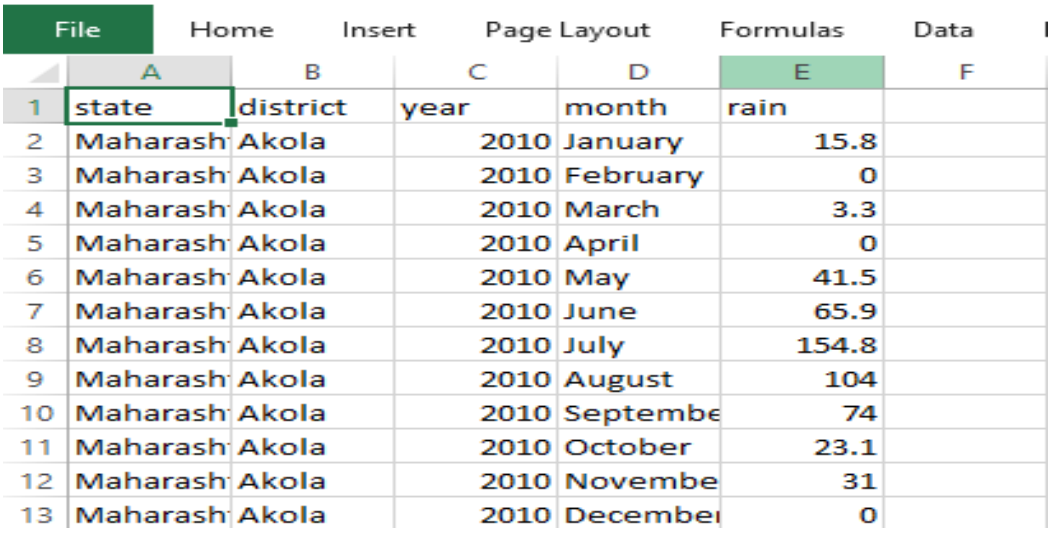


File Home Insert Page Layout Formulas Data

N	P	K	ph	ec
138	8.6	560	7.46	0.62
213	7.5	338	7.62	0.75
163	9.6	718	7.59	0.51
157	6.8	475	7.64	0.58
270	9.9	444	7.63	0.4
220	8.6	444	7.43	0.65
220	7.2	222	7.62	0.43
207	7	401	7.63	0.59
289	8.6	560	7.58	0.44

Figure 7: Proper Dataset Snap

The above image shows the data used as proper dataset which consist of the nutrient value for the classification of fertility level. There are total 880 soil sample have been gathered and based on that these data sets have been prepared. The above dataset have been devided into 80:20 for training and testing respectively.



File Home Insert Page Layout Formulas Data

	A	B	C	D	E	F
1	state	district	year	month	rain	
2	Maharash	Akola	2010	January	15.8	
3	Maharash	Akola	2010	February	0	
4	Maharash	Akola	2010	March	3.3	
5	Maharash	Akola	2010	April	0	
6	Maharash	Akola	2010	May	41.5	
7	Maharash	Akola	2010	June	65.9	
8	Maharash	Akola	2010	July	154.8	
9	Maharash	Akola	2010	August	104	
10	Maharash	Akola	2010	Septembe	74	
11	Maharash	Akola	2010	October	23.1	
12	Maharash	Akola	2010	Novembe	31	
13	Maharash	Akola	2010	Decembe	0	

Figure 8: Rainfall Dataset Snap

The above image is snap of the rainfall record of maharashtra region in different state. There are 3168 record have been used for the prediction and it also have been devided into 80:20 for training and testing. The rainfall data contains rainfall record month wise from year 2010 to 2017. It is used to predict the rainfall for the required year and month.

File Home Insert Page Layout Formulas

	A	B	C	D	E
1	crop	min_rainf	max_rainf	fertility	
2	Rice	500	1000	2	
3	Wheat	450	650	2	
4	Sorghum	450	650	1	
5	Maize	500	800	1	
6	Sugarcane	150	550	2	
7	Cotton	400	700	1	
8	Soybean	300	700	2	
9	Tomato	600	800	0	
10	Potato	500	700	2	
11	Onion	350	550	1	

Figure 9: Crop Database Snap

The above is snap of the crop database which records the list of crop, required minimum and maximum rainfall and fertility level of the soil in which it can be grown. The predicted crop is decided from this crop_db.csv which is important dataset for prediction.

4.2 Result

After running the project it start with evaluating datasets and preparing for the data. It shows the sample wise nutrientt value in soil.

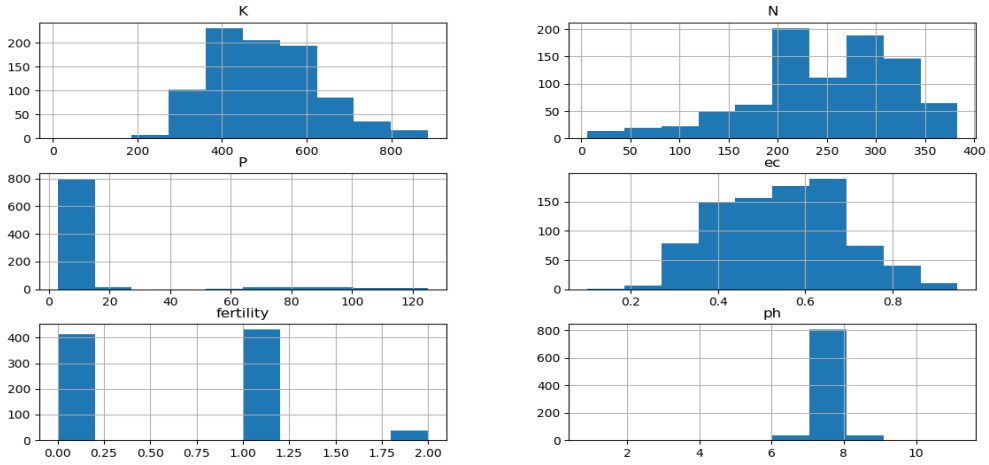


Figure 10: Soil sample dataset overview

The above figure shows nutrietion value present in soil sample in dataset.

Once the graphical representation of dataset is displayed it will the ask enduser or farmer to enter their feld nutrition value. Here it ask for the Potassium(K), Phosphorus(P), Nitrogen(N), Electrical Conductivity(ec), pH value(pH) of the soil, district, year and month in which farmer want to plough.

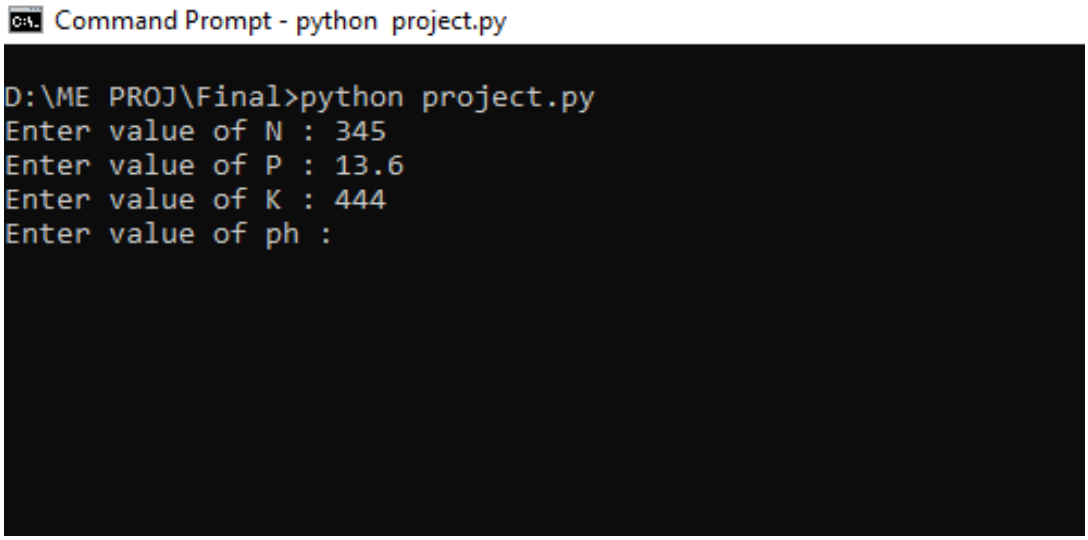


Figure 11: Input for the soil nutrient

After entering these data by enduser, it start with evaluating the result. First of all it presents the ROC-AUC Curve for Random Forest Classification for Training and Testing Dataset.

4.2.1 ROC curve

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

True Positive Rate

False Positive Rate

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

$$TPR=TP/(TP+FN)$$

False Positive Rate (**FPR**) is defined as follows:

$$FPR=FP/(FP+TN)$$

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

4.2.2 AUC: Area Under the ROC Curve

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

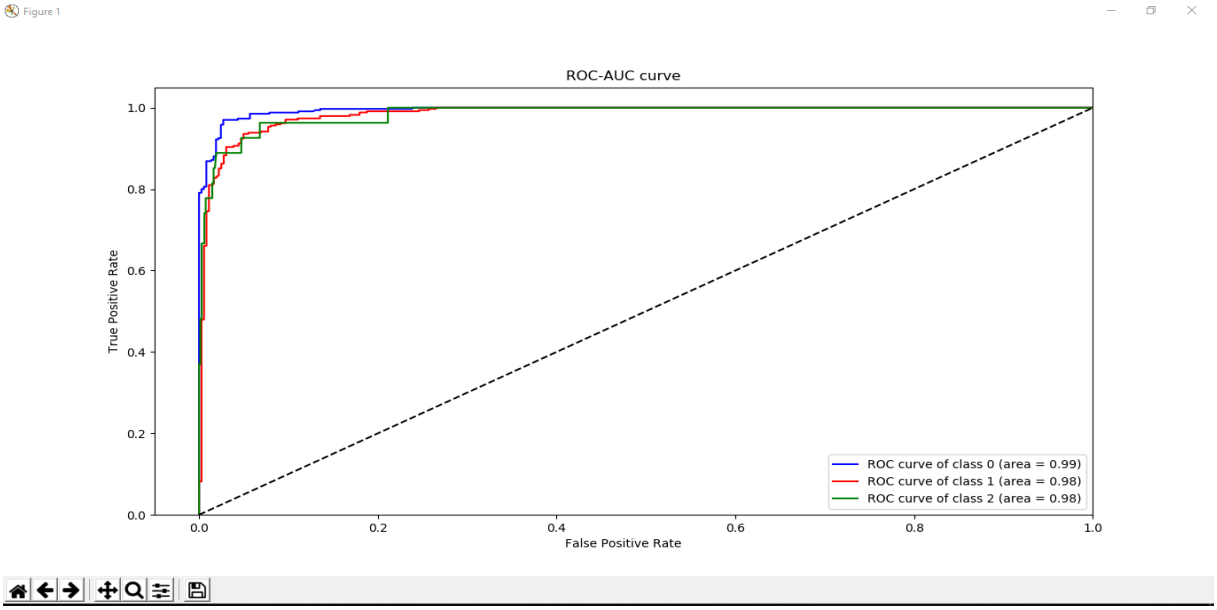


Figure 12: ROC Curve for the Training Dataset

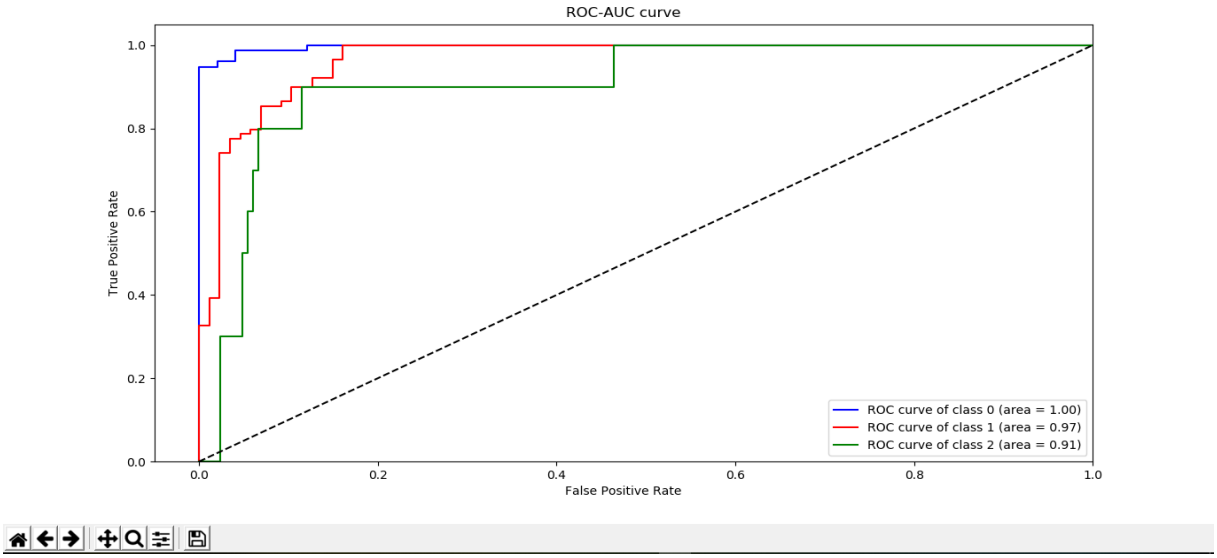


Figure 13: ROC Curve for the Testing Dataset

Along with displaying this ROC AUC Curve it displays other parameters for the algorithm. As shown in below image, this system displayed the performace based on different parameters.

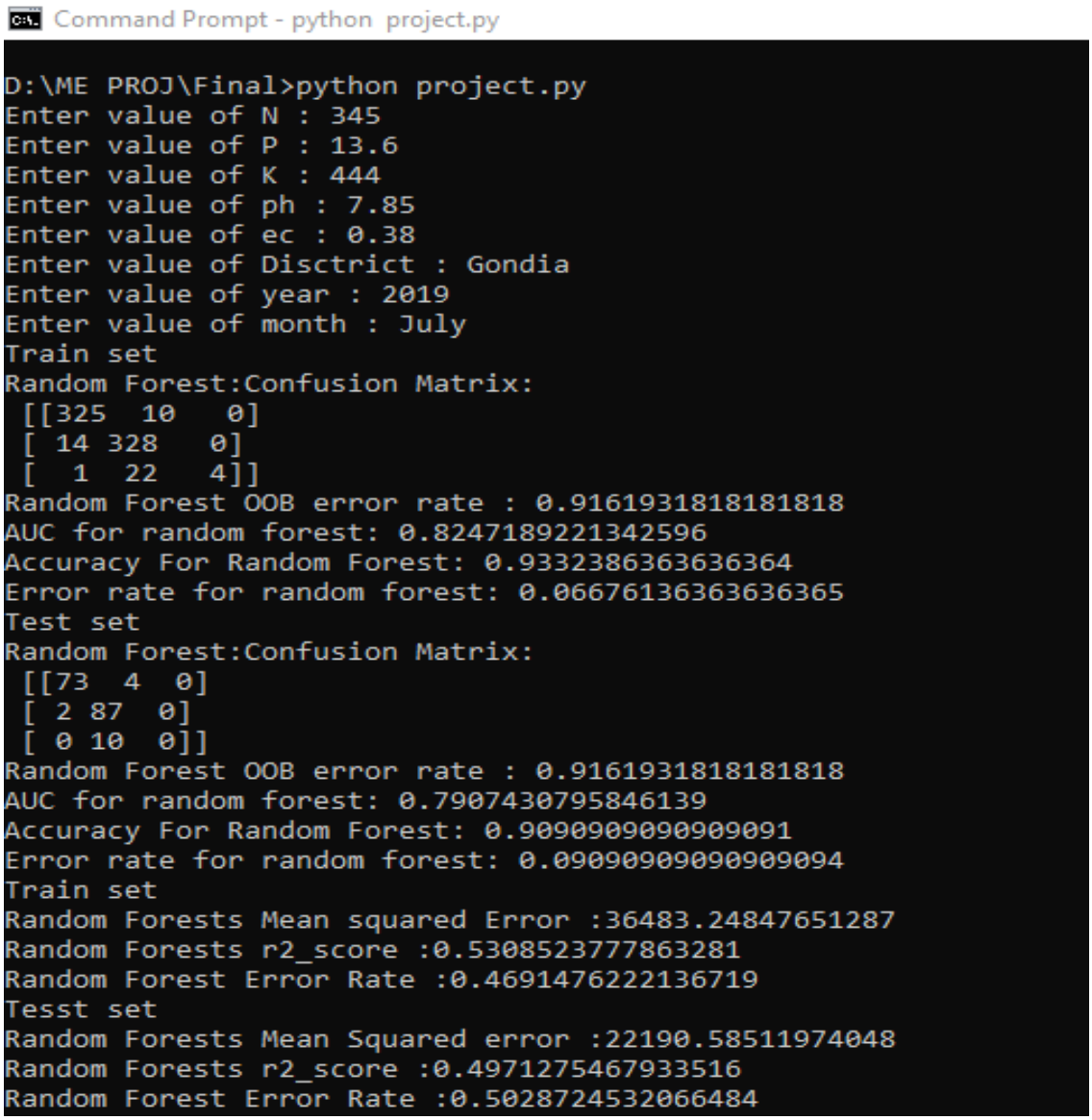


Figure 14: Algorithm Performance

The algorithm measure performance on different parameters. Few parameters like accuracy, OOB Error Rate, Confusion Matrix, Error Rate, Mean Squared Error, R^2 Score. Below we have given description for every parameter.

4.2.3 Accuracy:

Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It works well only if there are equal number of samples belonging to each class.

For example, consider that there are 98% samples of class A and 2% samples of class B in our training set. Then our model can easily get **98% training accuracy** by simply predicting every training sample belonging to class A.

When the same model is tested on a test set with 60% samples of class A and 40% samples of class B, then the **test accuracy would drop down to 60%.** Accuracy is great, but gives us the false sense of achieving high accuracy.

The real problem arises, when the cost of misclassification of the minor class samples are very high. If we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests.

4.2.4 Confusion Matrix:

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

Lets assume we have a binary classification problem. We have some samples belonging to two classes : YES or NO. Also, we have our own classifier which predicts a class for a given input sample. On testing our model on 165 samples ,we get the following result.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

There are 4 important terms :

1. **True Positives** : The cases in which we predicted YES and the actual output was also YES.
2. **True Negatives** : The cases in which we predicted NO and the actual output was NO.
3. **False Positives** : The cases in which we predicted YES and the actual output was NO.
4. **False Negatives** : The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the **"main diagonal"** i.e

$$Accuracy = \frac{TruePositives + FalseNegatives}{TotalNumberofSamples}$$

$$\therefore Accuracy = \frac{100 + 50}{165} = 0.91$$

Confusion Matrix forms the basis for the other types of metrics.

4.2.5 Area Under Curve:

Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Before defining AUC, let us understand two basic terms :

True Positive Rate (Sensitivity) : True Positive Rate is defined as *TP/ (FN*+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive}$$

False Positive Rate (Specificity) : False Positive Rate is defined as *FP / (*FP+TN). False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

$$FalsePositiveRate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

False Positive Rate and True Positive Rate both have values in the range [0, 1]. FPR and TPR bot hare computed at threshold values such as (0.00, 0.02, 0.04, …., 1.00) and a graph is drawn. AUC is the area under the curve of plot False Positive Rate vs True Positive Rate at different points in [0, 1].

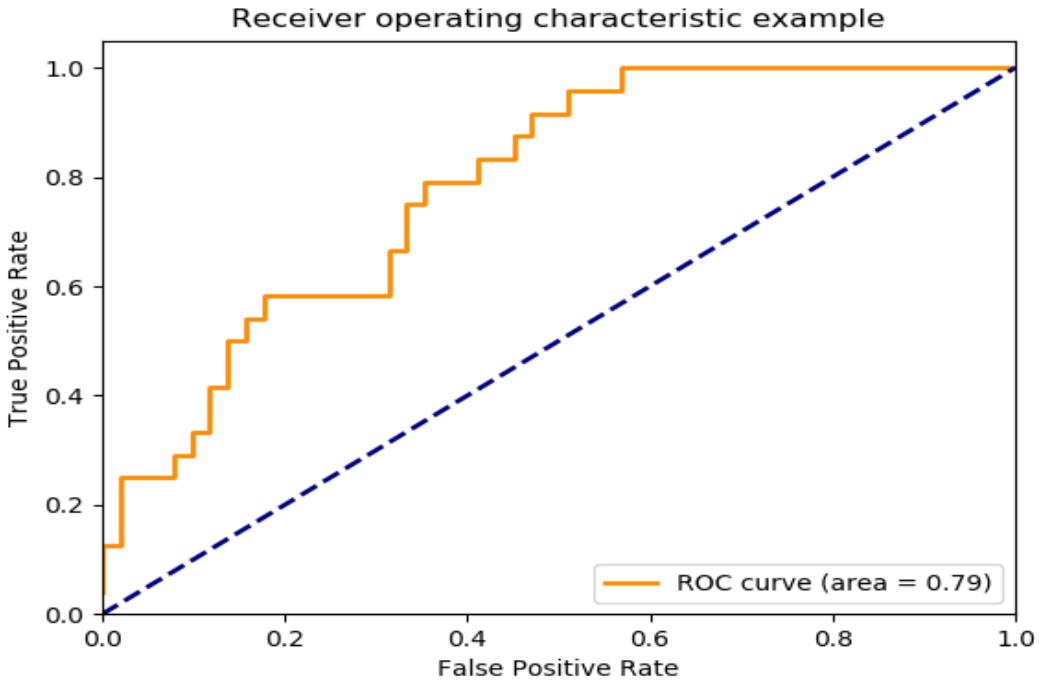


Figure 15: ROC Curve

As evident, AUC has a range of [0, 1]. The greater the value, the better is the performance of our model.

4.2.6 Mean Squared Error:

Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the **square** of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

4.2.7 Precision:

Precision is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy [27]. Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives. False positives are cases the model incorrectly labels as

positive that are actually negative, or in our example, individuals the model classifies as terrorists that are not. While recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant.[27]

$$Precision = TP/(TP+FP)$$

4.2.8 Recall:

Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned[27]. Recall is the ratio of correctly predicted positive observations to the all observations in actual.

$$Recall = TP/(TP+FN)$$

After dsplaying this performance measurement it displays predicted crop as shown below.

Figure 16: Predicted Crop

4.3 Result Obtained by Standard Random forest algorithm

As seen in below images, following results are obtained by using Traditional Random Forest algorithm. First of all it shows the dataset overview which is common for all.

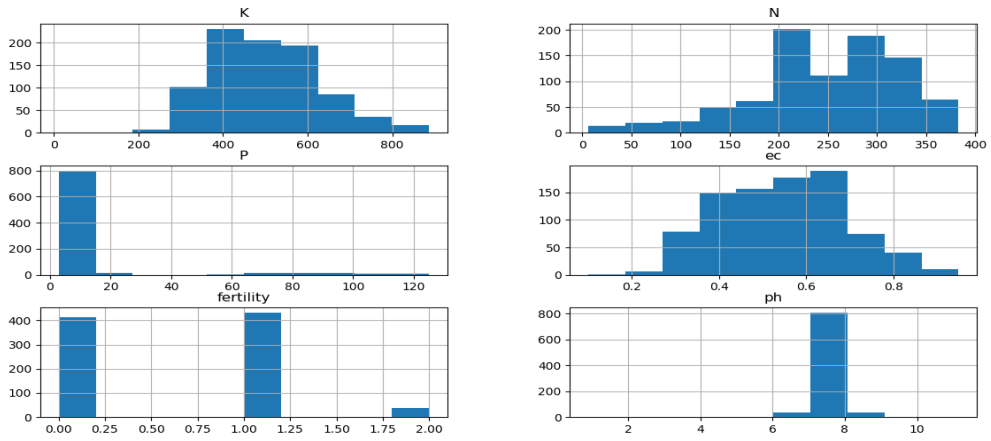


Figure 17: Soil sample dataset overview

After this it start with the performing performance analysis. First of all it diplays the ROC Curve for Training and Testing dataset.

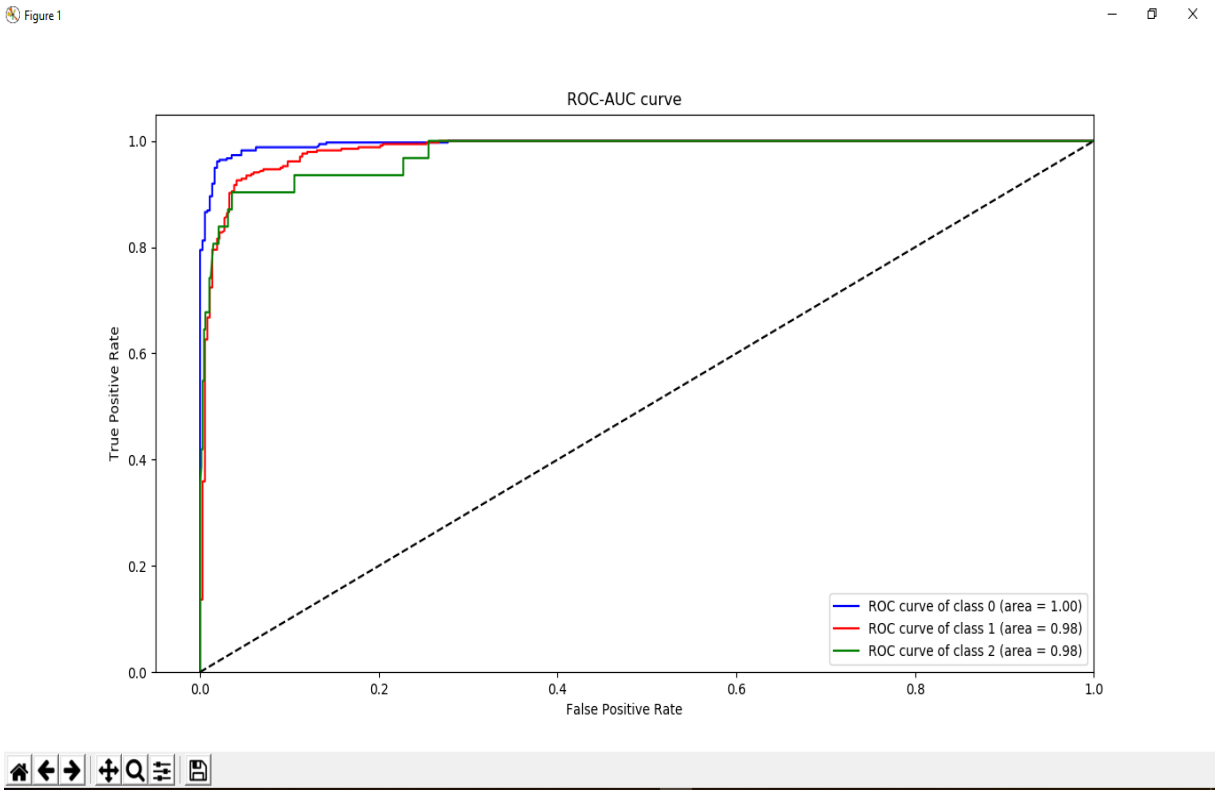


Figure 18: ROC Curve for the Training Dataset

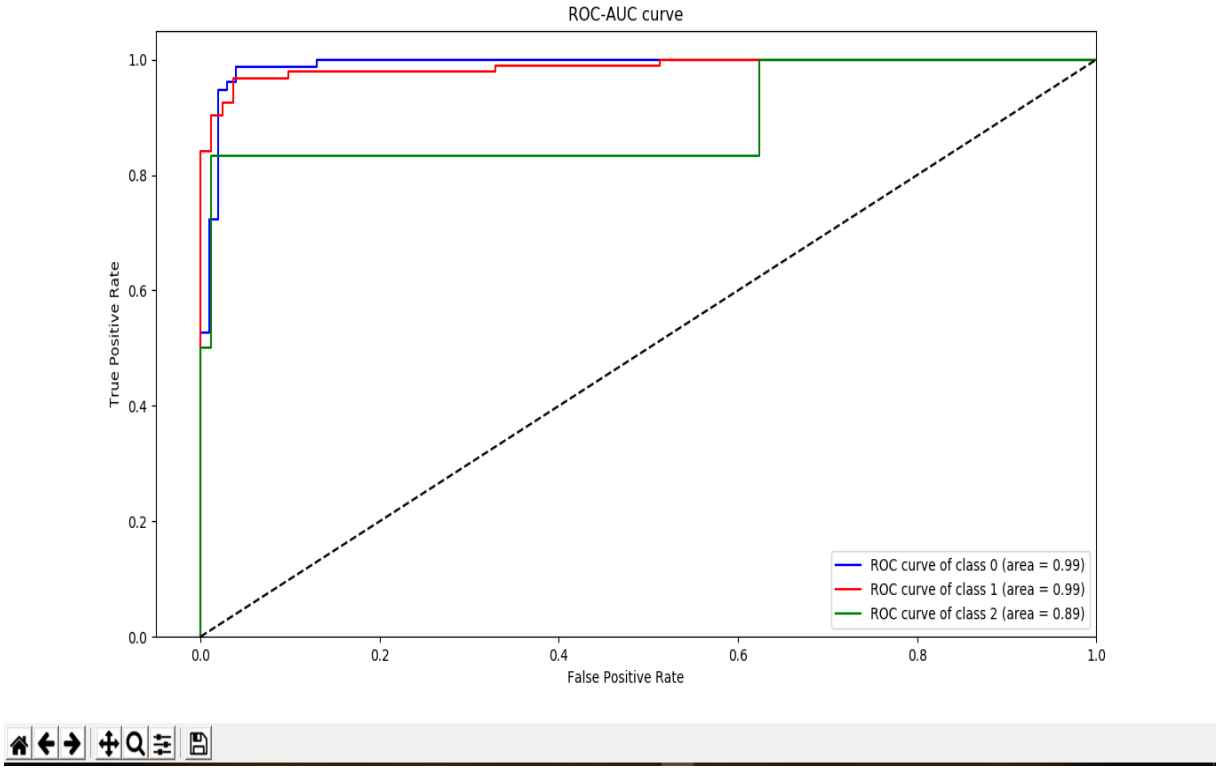


Figure 19: ROC Curve for the Testing Dataset

As shown in above figures, ROC curve for all class in training as well as testing dataset lies above 0.9 which shows that it is best fitting algorithm for classifying dataset. After displaying ROC Curve it analyses algorithm on different parameter for the data as shown in below image.

```
D:\ME PROJ\Final\Crop_Prediction_new\Crop_Prediction_16-7-2019>python rf_rf.py
Enter value of N : 333
Enter value of P : 7.5
Enter value of K : 507
Enter value of ph : 7.53
Enter value of ec : 0.54
Enter value of Disctrict : Pune
Enter value of year : 2019
Enter value of month : July
Train set
Random Forest:Confusion Matrix:
 [[315   7   0]
 [  7 343   0]
 [  0  30   2]]
Random Forest OOB error rate : 0.9034090909090909
AUC for random forest: 0.8163194147040125
Accuracy For Random Forest: 0.9375
Error rate for random forest: 0.0625
Classification Report :
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       322
           1       0.90      0.98      0.94       350
           2       1.00      0.06      0.12        32

    accuracy                           0.94       704
   macro avg       0.96      0.67      0.68       704
weighted avg       0.94      0.94      0.92       704

Test set
Random Forest:Confusion Matrix:
 [[89  1  0]
 [ 3 78  0]
 [ 1  4  0]]
Random Forest OOB error rate : 0.9034090909090909
AUC for random forest: 0.8087847741662512
Accuracy For Random Forest: 0.9488636363636364
Error rate for random forest: 0.051136363636363646
Classification Report :
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        90
           1       0.94      0.96      0.95        81
           2       0.00      0.00      0.00         5

    accuracy                           0.95       176
   macro avg       0.63      0.65      0.64       176
weighted avg       0.92      0.95      0.94       176

Train set
Random Forests Mean squared Error :12377.938172529062
Random Forests r2_score :0.6972169570792548
Random Forest Error Rate :0.3027830429207452
Tesst set
Random Forests Mean Squared error :16968.844106622622
Random Forests r2_score :0.6162273273647694
Random Forest Error Rate :0.3837726726352306
Predicted Values Using Random Forest as Classification and Random Forest Regression :

Predicted Fertility using random forest classification :
 [1]
Predicited rainfall using random forest regression :
 [309.19044908]
```

Figure 20: Algorithm Performance for Traditional Random Forest Algorithm

As shown in above snap,Random Forest Classification is analysed on OOB Error Rate, AUC, Accuracy and Error Rate where as Random Forest regression is analyzed over

Mean Squared Error,R^2 Score and Error rate. Following Table shows the performance measurement for the algorithm.

Table 1: Standard Random Forest Performance

Random Forest Classification								
ROC			**OOB Error Rate**	**AUC**	**Accuraacy**	**Error Rate**	**Precision**	**Recall**
Class 0	Class 1	Class 2						
0.99	0.99	0.89	0.91	80.07	93.18	0.07	0.92	0.95

Random Forest Regression		
Mean Squared Error	**R^2 Score**	**Error Rate**
19369.75	0.49	0.51

The predicted values also shown in above snap. Predicted crop will be displayed in another window as shown below.

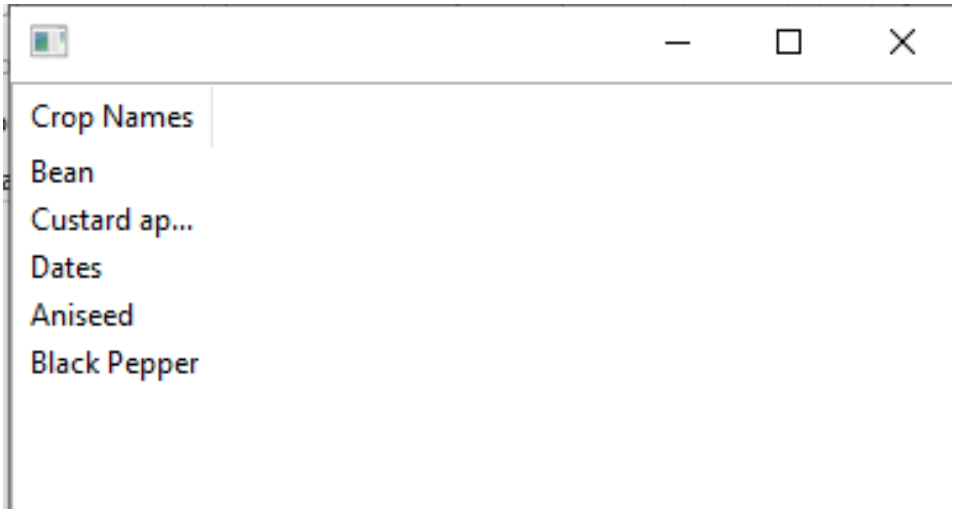


Figure 21: Predicted Crop by Standard RF

4.4 Result Obtained by ID3 algorithm:

As seen in below images, following results are obtained by using ID3 algorithm as classification as well as regression. First of all it shows the dataset overview which is common for all.

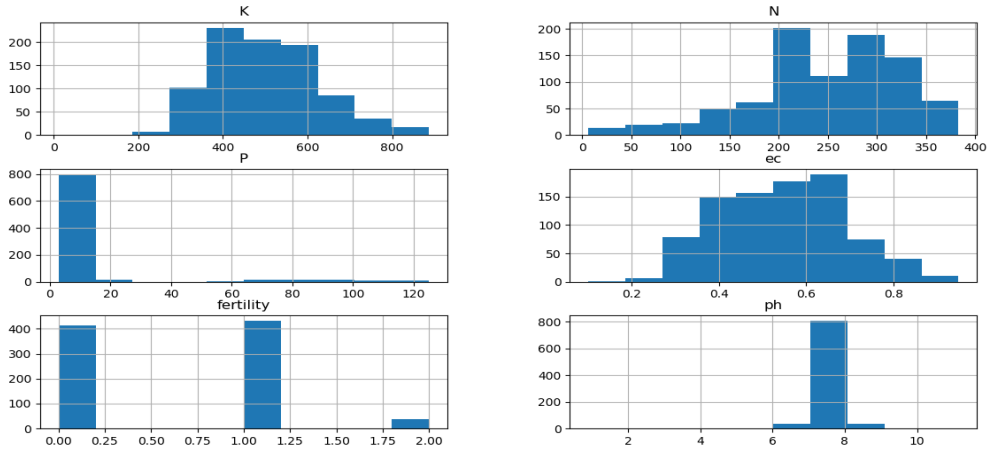


Figure 22: Soil Dataset Overview

After this it start with the performing performance analysis. First of all it diplays the ROC Curve for Training and Testing dataset.

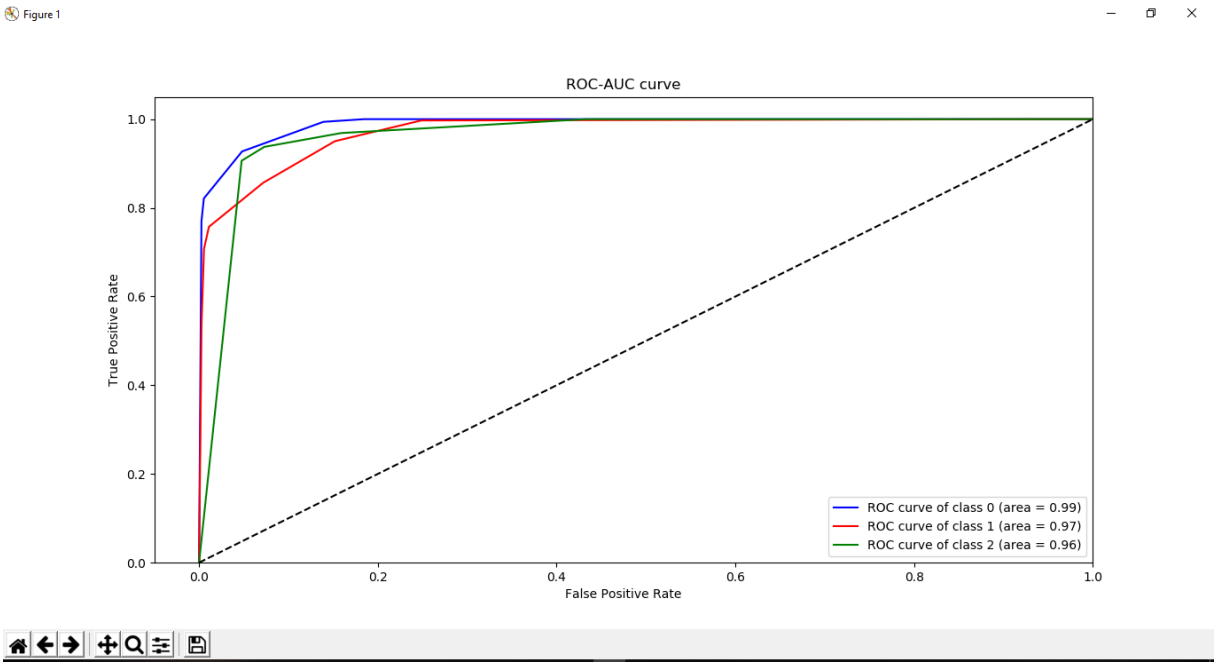


Figure 23: ROC Curve for the Training Dataset for ID3

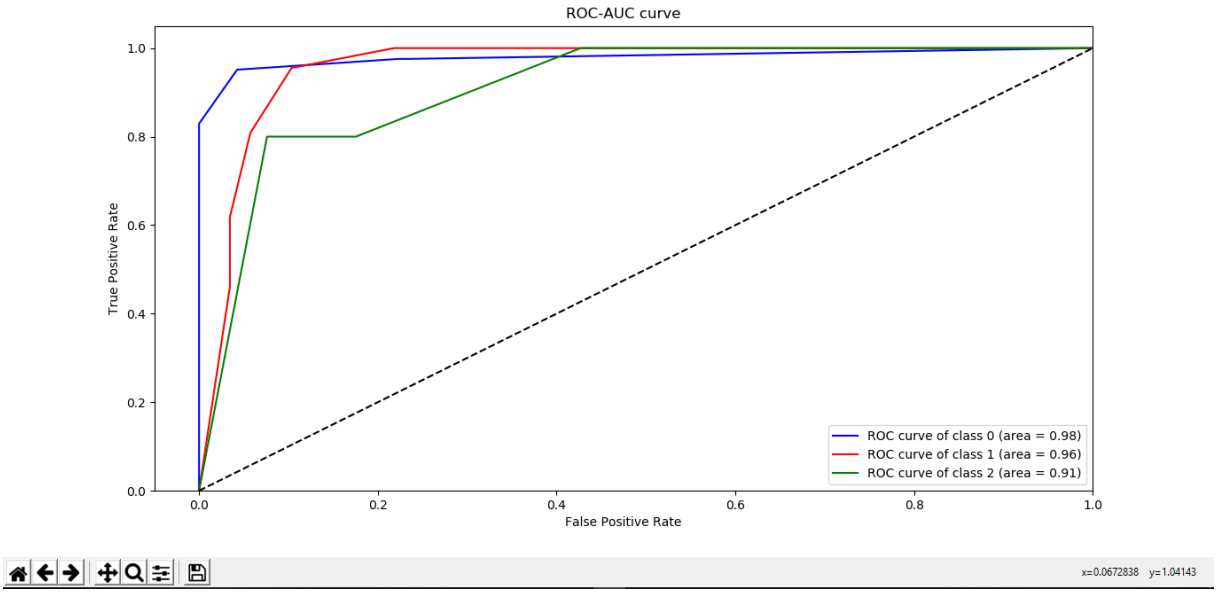


Figure 24: ROC Curve for the Training Dataset for ID3

As shown in above figures, ROC curve for all class in training as well as testing dataset lies above 0.9 which shows that it is better algorithm for classifying dataset but it is better than Random Forest. After displaying ROC Curve it analyses algorithm on different parameter for the data as shown in below image.

```
D:\ME PROJ\Final\Crop_Prediction_new\Crop_Prediction_16-7-2019>python ds_ds.py
Enter value of N : 333
Enter value of P : 7.5
Enter value of K : 507
Enter value of ph : 7.53
Enter value of ec : 0.54
Enter value of Disctrict : Pune
Enter value of year : 2019
Enter value of month : July
Decision_tree:train set
Decision_tree:Confusion Matrix:
  [[321  11   0]
 [ 12 327   0]
 [  1  32   0]]
AUC for Decision Tree: 0.7964524640089667
Decision_tree:Accuracy :  92.04545454545455
Error rate for Decision Tree: 0.07954545454545459
Classification Report :
```

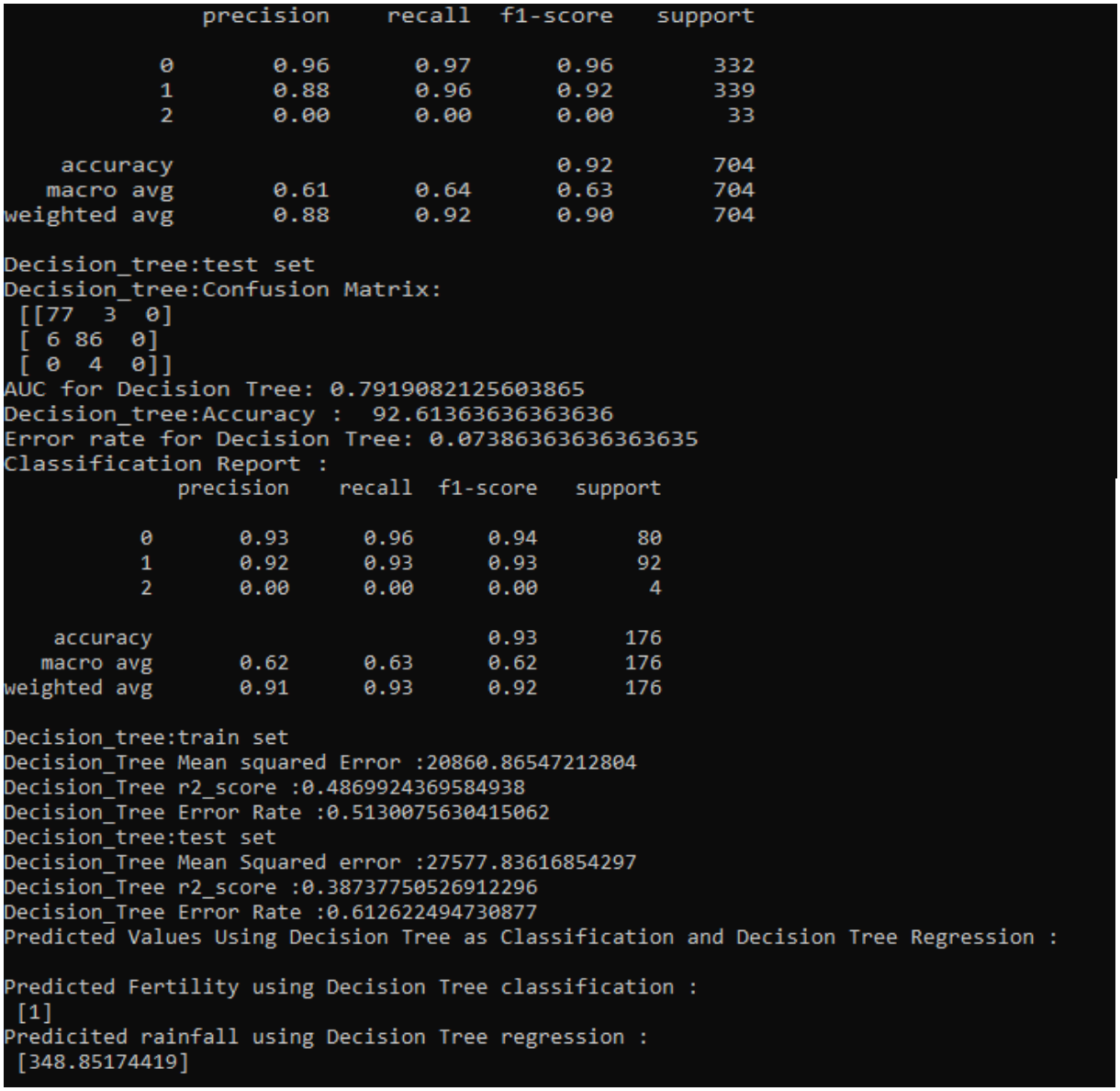


Figure 25: Performance Measurement for ID3

As shown in above snap,ID3 Classification is analysed on AUC, Accuracy and Error Rate where as Random Forest regression is analyzed over Mean Squared Error,R^2 Score and Error rate. Following Table shows the performance measurement for the algorithm.

Table 2: Performance table for ID3 as Classification and ID3 as Regression

ID3 as Classification							
ROC			**AUC**	**Accuraacy**	**Error Rate**	**Precision**	**Recall**
Class 0	Class 1	Class 2					
0.98	0.96	0.91	79.34	92.61	0.07	0.91	0.93

ID3 as Regression		
Mean Squared Error	**R^2 Score**	**Error Rate**
25850.17	0.45	0.55

The predicted values also shown in above snap. Predicted crop will be displayed in another window as shown below.

Figure 26: Predicted crop using ID3

4.5 Result Obtained by Enhanced Random Forest algorithm:

As seen in below images, following results are obtained by using Enhanced RF. First of all it shows the dataset overview which is common for all.

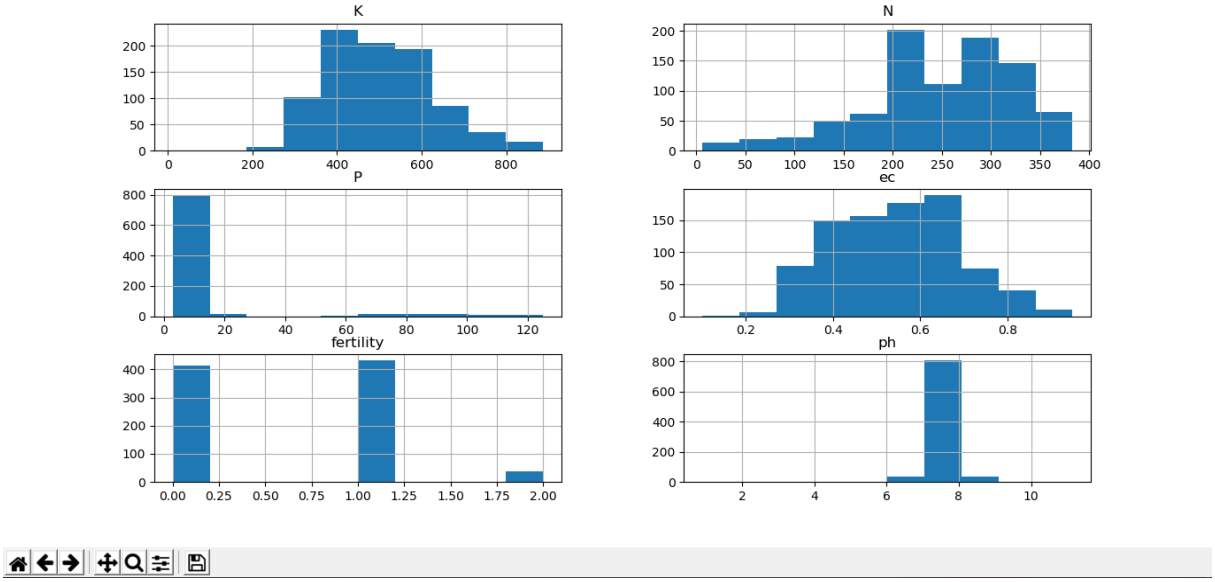


Figure 27: Soil Dataset Overview

After this it start with the performing performance analysis. First of all it diplays the ROC Curve for Training and Testing dataset.

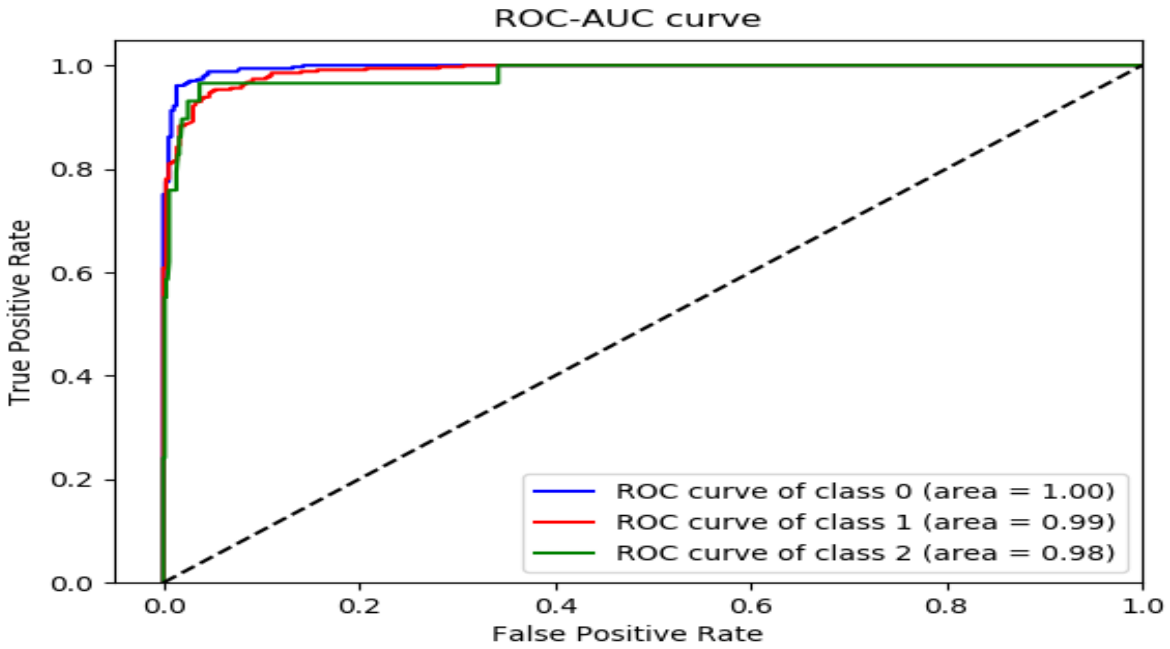


Figure 28: ROC Curve for the Training Dataset for Enhanced RF

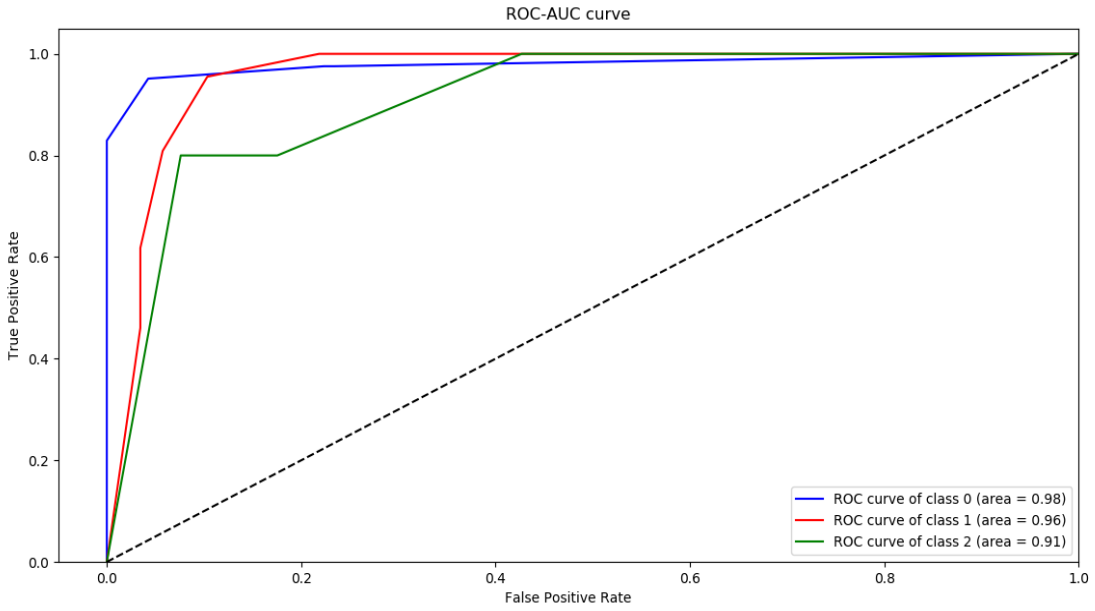


Figure 29: ROC Curve for the Training Dataset for Enhanced RF

As shown in above figures, ROC curve for all class in training as well as testing dataset lies above 0.9 which shows that it is better algorithm for classifying dataset but it is better than Random Forest. After displaying ROC Curve it analyses algorithm on different parameter for the data as shown in below image.

```
Enter value of N : 333
Enter value of P : 7.5
Enter value of K : 507
Enter value of ph : 7.53
Enter value of ec : 0.54
Enter value of Disctrict : Pune
Enter value of year : 2019
Enter value of month : July
Train set
Random Forest:Confusion Matrix:
 [[323  15   0]
 [  8 332   0]
 [  0  25   1]]
Random Forest OOB error rate : 0.9147727272727273
AUC for random forest: 0.8064675658474338
Accuracy For Random Forest: 0.9318181818181818
Error rate for random forest: 0.06818181818181823
Classification Report :
              precision    recall  f1-score   support

           0       0.98      0.96      0.97       338
           1       0.89      0.98      0.93       340
           2       1.00      0.04      0.07        26

    accuracy                           0.93       704
   macro avg       0.96      0.66      0.66       704
weighted avg       0.94      0.93      0.92       704

Test set
Random Forest:Confusion Matrix:
 [[68  6  0]
 [ 1 90  0]
 [ 1 10  0]]
Random Forest OOB error rate : 0.9147727272727273
AUC for random forest: 0.7833477951125011
Accuracy For Random Forest: 0.8977272727272727
Error rate for random forest: 0.10227272727272729
Classification Report :
              precision    recall  f1-score   support

           0       0.97      0.92      0.94        74
           1       0.85      0.99      0.91        91
           2       0.00      0.00      0.00        11

    accuracy                           0.90       176
   macro avg       0.61      0.64      0.62       176
weighted avg       0.85      0.90      0.87       176

Train set
Random Forests Mean squared Error :3377.3378007843435
Random Forests r2_score :0.9190338863701428
Random Forest Error Rate :0.08096611362985717
Tesst set
Random Forests Mean Squared error :4818.4080367691495
Random Forests r2_score :0.8820606870012592
Random Forest Error Rate :0.11793931299874083
Predicted Values Using Random Forest as Classification and Random Forest Regression :

Predicted Fertility using random forest classification :
 [1]
Predicited rainfall using random forest regression :
 [264.7439477]

D:\ME PROJ\Final\Crop_Prediction_new\Crop_Prediction_16-7-2019>
```

Figure 30: Performance Measurement for Enhanced RF

As shown in above snap,Enhanced RF Classification is analysed on AUC, Accuracy and Error Rate where as Enhanced Random Forest regression is analyzed over Mean Squared Error,R^2 Score and Error rate. Following Table shows the performance measurement for the algorithm.

Table 3: Performance table for ID3 as Classification and ID3 as Regression

Enhanced RF as Classification							
ROC			**AUC**	**Accuraacy**	**Error Rate**	**Precision**	**Recall**
Class 0	Class 1	Class 2					
1.00	0.99	0.98	79.53	93.18	0.07	0.97	0.92

Enhanced RF as Regression		
Mean Squared Error	**R^2 Score**	**Error Rate**
9336.36	0.68	0.31

The predicted values also shown in above snap. Predicted crop will be displayed in another window as shown below.

Figure 31: Predicted crop using Enhanced RF

4.6 Comparison

Table 4: Comparison of RF and ID3 against Enhanced RF

Regression	Enhanced Random Forest	Error Rate	0.25	0.21	0.23	0.23	0.21	0.22	0.2	0.24	0.21	0.23
	ID3		0.52	0.65	0.49	0.49	0.54	0.51	0.57	0.51	0.49	0.52
	Standard Random Forest		0.51	0.48	0.5	0.52	0.57	0.47	0.49	0.54	0.51	0.48
	Enhanced Random Forest	R2 Score	68.32	70.42	71.64	72.48	65.56	78.56	78.48	68.45	67.58	72.65
	ID3		50.26	48.46	49.64	48.65	47.35	49.85	47.67	51.65	52.26	48.65
	Standard Random Forest		49.15	50.12	50.1	48.16	50.16	50.75	49.26	46.82	50.1	51.46
	Enhanced Random Forest	Mean Squared Error	9336.36	7645.56	7546.3	8343.16	9435.23	7653.64	8323.8	8542.35	9356.64	7564.65
	ID3		19475.35	20365.95	18356.24	17356.32	19648.65	20315.94	17546.65	18356.26	17365.32	18653.23
	Standard Random Forest		19369.76	15647.23	17534.35	16547.24	13247.23	16326.23	17453.12	17354.32	18975.64	19346.64
	Enhanced Random Forest	Recall	0.94	0.95	0.94	0.94	0.94	0.94	0.95	0.95	0.94	0.94
	ID3		0.9	0.9	0.89	0.88	0.89	0.9	0.89	0.89	0.9	0.9
	Standard Random Forest		0.91	0.92	0.92	0.93	0.92	0.91	0.92	0.91	0.92	0.91
	Enhanced Random Forest	Precision	0.97	0.96	0.97	0.96	0.95	0.97	0.96	0.96	0.97	0.95
	ID3		0.9	0.91	0.9	0.9	0.92	0.91	0.9	0.91	0.91	0.92
	Standard Random Forest		0.93	0.95	0.94	0.95	0.94	0.94	0.95	0.94	0.95	0.93
Classification	Enhanced Random Forest	Error Rate	0.07	0.08	0.08	0.07	0.08	0.08	0.08	0.08	0.07	0.08
	ID3		0.08	0.09	0.09	0.08	0.09	0.1	0.09	0.09	0.1	0.09
	Standard Random Forest		0.09	0.09	0.08	0.08	0.09	0.09	0.1	0.09	0.08	0.08
	Enhanced Random Forest	Accuracy	93.2	93.4	92.3	92.8	92.2	93.4	92.1	92.2	92.4	92.8
	ID3		89.3	90.2	91.5	91.2	89.5	89.6	88.7	90.1	90.2	89.2
	Standard Random Forest		90.3	91.2	91.7	91.9	90.7	90.6	91.2	90.5	89.3	91.8
	Enhanced Random Forest	AUC	0.81	0.8	1	0.9	0.97	0.98	0.9	0.91	0.97	0.94
	ID3		0.72	0.76	0.8	0.93	0.94	0.72	0.94	0.9	0.81	0.85
	Standard Random Forest		0.79	0.78	0.9	0.99	0.95	0.79	1	0.89	0.88	0.86
	Enhanced Random Forest	OOB Error rate	0.9	0.9	0.89	0.9	0.88	0.87	0.88	0.85	0.87	0.88
	ID3		0.91	0.9	0.9	0.91	0.92	0.89	0.9	0.88	0.91	0.9
	Standard Random Forest		0.91	0.89	0.92	0.88	0.92	0.9	0.87	0.92	0.9	0.89
			Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10

As compared in above table, Random forest classification algorithm gives higher ROC Value, Less OOB Error Rate, higher AUC, Accuracy and less Error Rate against Random Forest Regression in comparison to ID3 algorithm as regression. Even ID3 classification didn't perform well neither against Random Forest nor ID3 Regression

algorithm. Also, in Regression Comparison table, Random Forest Regression perform very well. It returns adequate lower Mean Squared Error compared to ID3 regression against Random Forest algorithm as classification. Also it returns with the higher R^2 Score and lower Error rate in compared to ID3 regression algorithm against any other classification algorithm.

4.6.1 Accuracy Comparison:

In below graph chart accuracy of ID3, Standard Random Forest and Enhanced Random Forest is compared and Enhanced Random Forest gives best accuracy while Standard Random Forest have comparatively less accuracy for all dataset.

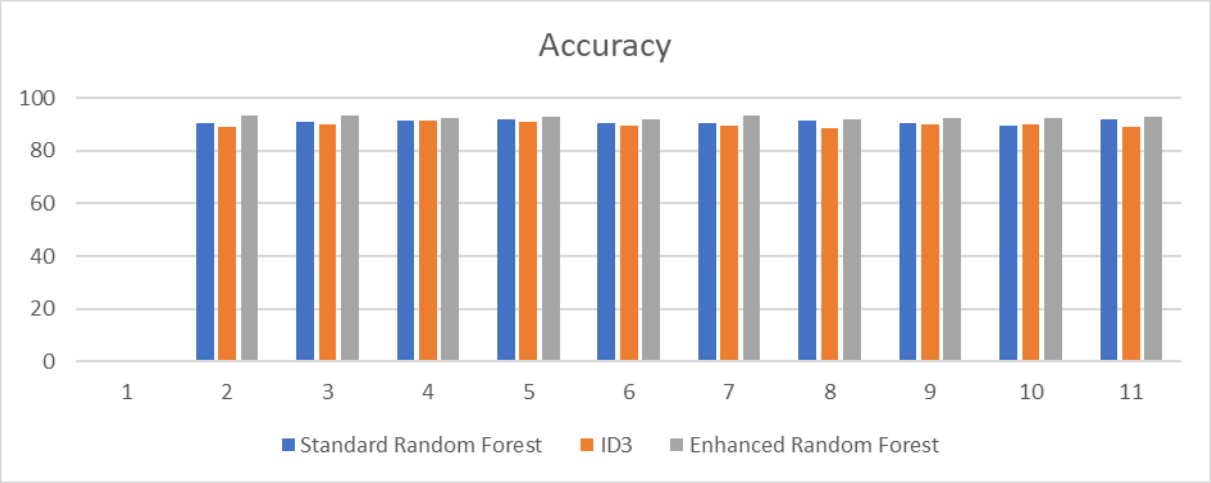


Figure 32: Accuracy Comparison

4.6.2 Error Rate Comparison:

In below graph chart error rate of ID3, Standard Random Forest and Enhanced Random Forest is compared and Enhanced Random Forest gives lowest error rate while Standard Random Forest have comparatively higher error rate for all dataset.

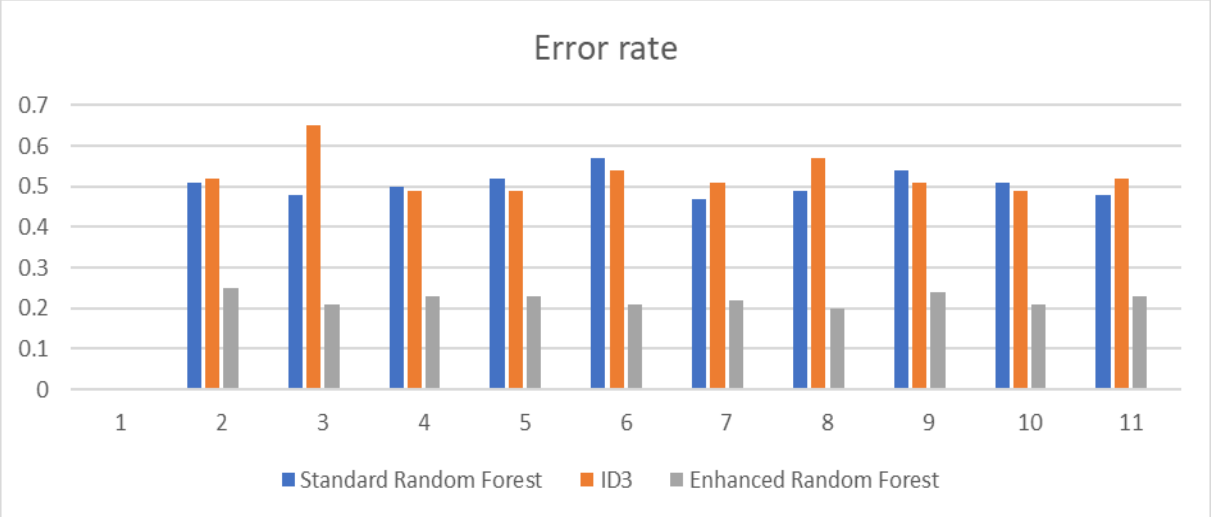


Figure 33: Error Rate Comparison

4.6.3 R^2 Score:

In below graph chart R^2 Score of ID3, Standard Random Forest and Enhanced Random Forest is compared and Enhanced Random Forest gives highest score while Standard Random Forest have comparatively less score for all dataset.

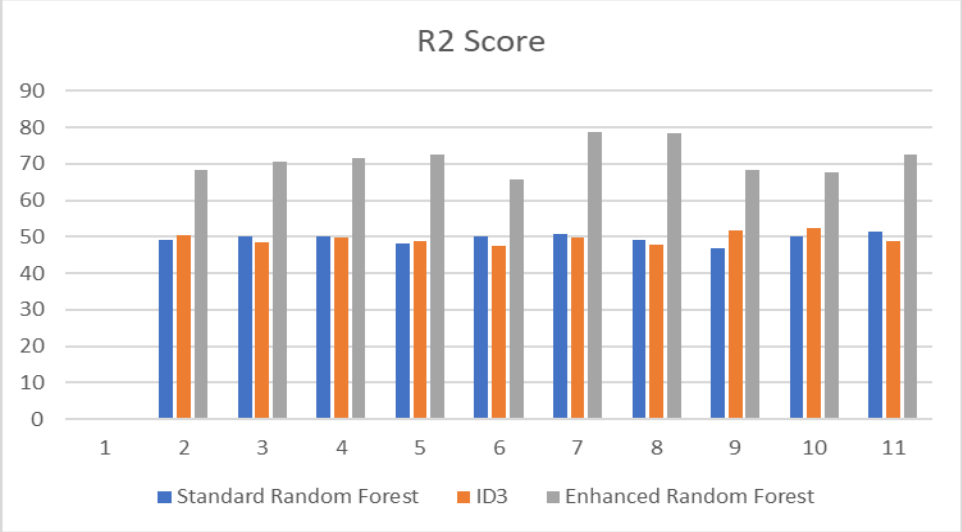


Figure 34: R^2 Comparison

4.6.4 Precision:

In below graph chart Precision of ID3, Standard Random Forest and Enhanced Random Forest is compared and Enhanced Random Forest gives highest Precision while Standard Random Forest have comparatively less Precision for all dataset.

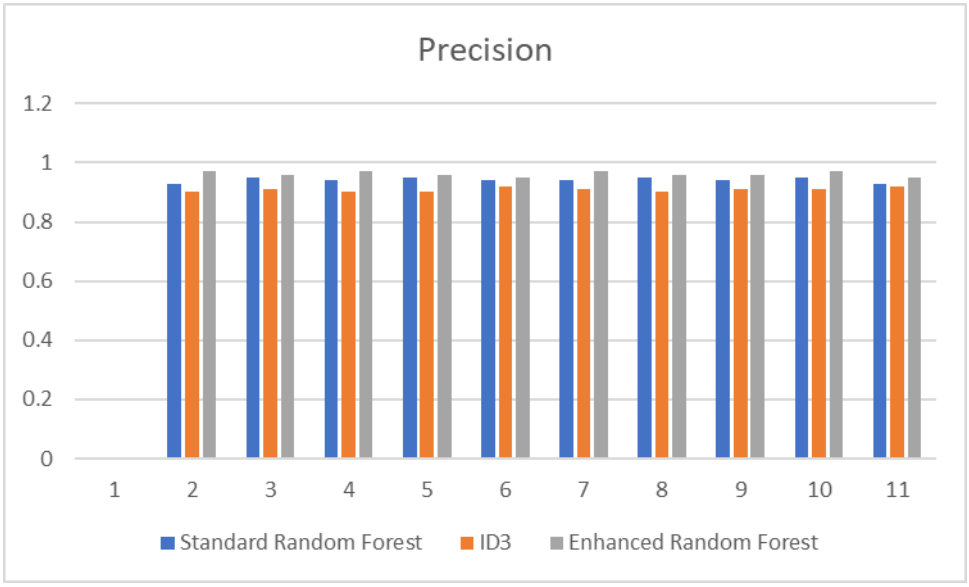


Figure 35: Precision Comparison

CHAPTER 5. Conclusion

Conclusion

Earlier yield production was decided based on farmers experience where technology involvement was not there which gives accurate answer to decide the crop to plough. Therefore, in order to help farmers to decide the crop to plough for their financial as well as social benefits crop prediction system make use of Random Forest as classification as well as regression. Classification algorithm classifies the soil sample based on the available nutrient in soil into different class of soil where as regression predicts the expected rainfall for the entered year and month in which farmer want to plough.

Enhanced Random Forest classification and regression which performed in comparison. The classification comparison is based on the parameter ROC Curve, AUC, OOB Error Rate, Accuracy and Error Rate, where as Regression comparison is based on parameter Mean Squared error, R^2 Score and Error rate.

The planned model work presents comparison of Random forest Classification combined with Random Forest Regression and ID3 Regression and Enhanced Random Forest. Different soil sample have been used to compare the algorithm and it is concluded that Enhanced Random Forest as classification and Regression performed better in term of RUC Curve, AUC, Accuracy, Error Rate and OOB Error Rate. Accuracy is most important parameter that demonstrate the performance of any algorithm. It is observed that accuracy of Enhanced Random Forest as Classification and Regression combined is better in classifying the dataset and predicting the result.

Future Scope

The future of the Random Forest as classification and Regression involves predicting the pesticides and fertilisers to be used to improve fertility level of soil based on current micro and macro nutrient available in soil. Random Forest as classification and Regression is also helpful in predicting the rainfall for coming years based on previous rainfall trend.

References

Literature Cited

[1] Supriya D M "Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining Approach" in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 5, May 2017.

[2] Vaneesbeer Singh, Abid Sarwar "Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach" in International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May – June 2017.

[3] Profile of Maharashtra and selected districts, shodhganga.inflibnet.ac.in/bitstream/10603/121515/13/13_chapter4.pdf

[4] Andrew.W "Moore Professor School of Computer Science Carnegie Mellon University", Naïve Bayes Classifiers, www.cs.cmu.edu/~awm awm@cs.cmu.edu

[5] Andrew.W "Moore Professor School of Computer Science Carnegie Mellon University", Naïve Bayes Classifiers, www.cs.cmu.edu/~awm awm@cs.cmu.edu

[6] B. Bhattacharya, D.P. Solomatine "Machine learning in soil classification" Elsevier 2006 Special Issue, Neural Networks 19 (2006) 186–195

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio "Generative Adversarial Nets"

[8] Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems. **52**: 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377

[9] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering" (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284&ndash, 292. ISBN 0-521-64298-1. MR 2012999.

[10] Coates, Adam; Ng, Andrew Y. (2012). "Learning feature representations with k-means" (PDF). In G. Montavon, G. B. Orr, K.-R. Müller. *Neural Networks: Tricks of the Trade*. Springer.

[11] Csurka, Gabriella; Dance, Christopher C.; Fan, Lixin; Willamowski, Jutta; Bray, Cédric (2004). *Visual categorization with bags of keypoints* (PDF). ECCV Workshop on Statistical Learning in Computer Vision.

[12] Coates, Adam; Lee, Honglak; Ng, Andrew Y. (2011). *An analysis of single-layer networks in unsupervised feature learning* (PDF). International Conference on Artificial Intelligence and Statistics (AISTATS). Archived from the original (PDF) on 2013-05-10.

[13] Schwenker, Friedhelm; Kestler, Hans A.; Palm, Günther (2001). "Three learning phases for radial-basis-function networks". *Neural Networks*. **14** (4–5): 439–458. CiteSeerX 10.1.1.109.312. doi:10.1016/s0893-6080(01)00027-2

[14] Geetha MCS. Implementation of association rule mining for different soil types in agriculture. International Journal of Advanced Research in Computer and Communication Engineering. 2015 Apr; 4(4):520–2.

[15] Knowledge Discovery and Data Mining to Identify Agricultural Patterns, Kulwant Kaur, Maninderpal Singh, IJESRT [1337-1345], March, 2014

[16] G.Kesavaraj, Dr.S.Sukumaran *"*A Study on Classification Techniques in Data Mining" IEEE – 31661

[17] Angiulli, F., Basta, S. and Pizzuti, C. 2006. *Distance-based detection and prediction of outliers.* IEEE Transactions on Knowledge and Data Engineering, **18:**145- 160.

[18] Bader-El-Den, M. and Gaber, M. 2012. Garf: Towards self-optimised random forests. In T. Huang, Z.Zeng, C. Li, & C.-S. Leung (Eds.), *Neural Information Processing – 19th International Conference, ICONIP 2012, Doha, Qatar, Proceedings, Part II*, Lecture Notes in ComputerScience. Berlin: Springer. pp. 506–515.

[19] Banfield, R.E., Hall, L.O., Bowyer, K.W. and Kegelmeyer, W.P. 2006. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **29(1)**: 173–180.

[20] Barnett ,V and Lewis, T. 1978. *Outliers in statistical data,* John Wiley & Sons. 1978. p.1.

[21] Bernard, S., Heutte, L., Adam, S. 2009. *On the selection of decision trees in random forests*. International Joint Conference on Neural Network , pp. 302–307.

[22] Biau, G., Devroye, L., and Lugosi, G., 2008. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research,* **9**: 2015–2033.

[23] Biau, G. 2012. Analysis of a random forests model. *Journal of Machine Learning Research*, **13**:1063–1095.

[24] Breiman, L. 2004. Consistency for a simple model of random forests. Technical report, University of California at Berkeley.

[25] Cao, H., Zhou, Y., Shou, L. and Chen, G. 2010. Attribute outlier detection over data streams. In: *Proceedings of the 15th international conference on Database Systems for Advanced Applications - Volume Part II*. DASFAA'10 (Springer-Verlag, Berlin, Heidelberg). pp 16–230.

[26] Criminisi, A. and Shotton, J. 2013. Decision Forests for omputer Vision and Medical Image Analysis. Springer London. XIX, pp 1-368. ISBN 978-1-4471-4929-3.

[27] https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[28] Yang F., Lu W., Luo L., Li T., Margin optimization based pruning for random forest, *Neurocomputing*, 94, pages 54–63 (2012)

[29] Schapire R. E., The Boosting Approach to Machine Learning an Overview, *Nonlinear Estimation and Classification*, LNS Volume 171, pages 149-171, Springer Science plus Business Media New York, 2003

[30] Menze B. H., Kelm B. M., Splitthoff D. N., Koethe U., and Hamprecht F. A., On Oblique Random Forests, Proceedings of ECML PKDD, Part II, LNAI 6912, pages 453-469,Springer-Verlag (2011)

[31] http:// code.google.com/p/parf

[32] http://www.cs.waikato.ac.nz/ml/weka

[33] http://cran.r-project.org

[34] http://sci2s.ugr.es/keel

[35] www.randomjungle.de

[36] http://hadoop.apache.org

[37] www.kdd.org

[38] Soil Data Set- http://soilhealth.dac.gov.in/PublicReports/NutrientStatusFarmerWise

Publication

Journal papers

Sr. No.	Paper Title	Publication details	Phase of project
1	CROP PREDICTION BY SOIL ANALYSIS USING ENHANCED RANDOM FOREST ALGORITHM FOR VIDARBHA REGIONS	International Conference on Intelligent Systems and Communication Networks (IC-ISCN 2019)	Proposed Idea Published
2	COMPARATIVE STUDY ON ENHANCED RANDOM FOREST ALGORITHM BY CROP PREDICTION	International Journal for Research in Engineering Application & Management (IJREAM)	Result and Comparative Study Published

Crop Prediction by Soil Analysis using Enhanced Random Forest Algorithm for Vidarbha Regions

Priyankar Ravindra Tiwari, Prof. Anand Khandare

Department of Computer Engineering,
Thakur College of Engineering and Technology,
Mumbai,India
priyankartiwarimy@gmail.com, anand.khandare1983@gmail.com

***Abstract*—In the years since its independence, India has made immense progress towards food security. Indian population has tripled, and food-grain production more than quadrupled. India ranked in the world's five largest producers of over 80% of agricultural produce items. India exported $39 billion worth of agricultural products in 2013, making it the seventh largest agricultural exporter worldwide, and the sixth largest net exporter. This represents explosive growth, as in 2004 net export were about $5 billion. India is the fastest growing exporter of agricultural products over a 10-year period, its $39 billion of net exports is more than double the combined exports of the European Union. So, yield prediction is very popular among farmers these days. Earlier yield prediction was performed by considering the farmer's experience on a particular field, weather and crop. Since farmers don't have knowledge about the presence of the nutrients and they don't have the idea about the crop to plough and pesticides to be used due to which performance of agriculture is degrading in economy and farmers getting into loss which they have to pay from their own pocket, which is also a root cause of farmers suicide. This makes the problem of predicting the yielding of crops an interesting challenge. Proposed system uses machine learning techniques in order to predict the category of the analyzed soil datasets. The category, thus predicted will indicate the yielding of crops. The problem of predicting the crop yield is formalized as a classification rule, where Boosted Tree and Random Forest algorithms will be used and for feature extraction K-Means will be used.**

Keywords—Random Forest, K-Means, Boosted Tree, Crop, Prediction, Vidarbha.

I. INTRODUCTION (*HEADING 1*)

Agriculture is the backbone of the Indian. The agriculture data increases day by day. Since a large population lives in rural areas and is directly or indirectly dependent on agriculture for a living. Outlay from farming forms the main source for the farming community[1]. The essential requirements for harvesting are water resources and ability to buy seeds, fertilizers, pesticides, labour etc. Most farmers raise the required capital by compromising on other essential expenditures, and when it is still insufficient they resort to credit from sources like banks and private commercial institutions. In such a situation, the repayment is dependent on the success of the harvest. If the harvest fails even once due to several factors, like bad weather pattern; soil type; improper, excessive, and ill-timed application of both fertilizers and pesticides; adulterated seeds and pesticides etc. Most power of soil in nature comes from soil survey efforts. Soil survey, or soil mapping, is the process of determining the soil types or other holding of the soil cover over a landscape, and mapping them for others to understand and use. Primary data for the soil survey is acquired by area sampling and supported by remote sensing. As the volume of data increase, it requires involuntary way for these data to be extracted when needed. Machine Learning can be used for pretend the next trends of agricultural processes. Every soil is a mixture of three main components: sand, clay and silt. Based on these factors we want to predict the soil for a particular cultivation.

Machine Learning task can be classified into two categories: Descriptive Machine Learning and Predictive Machine Learning. Descriptive Machine Learning tasks qualify the general properties of the database while predictive Machine Learning is used to predict expressed values based on patterns determined from known results.

Classification and prediction are two forms of data analysis that can be used to solution models describing important data classes or to predict future data trends. It is a process in which models learn to pretend a class label from a set of training data which can then be used to predict discrete class labels on new sample. Different subsets of these parameters are used in different prediction models for different crops.

II. LITERATURE SURVEY

The amount of data doubles almost every year. Hence, there is an urgent need for a new generation of computationally intelligent techniques and tools to assist human beings in extracting useful information from the rapidly growing volume of data. It is realized that extracting knowledge from large amount of data is typically ill-defined and it is difficult to model with large scale solution spaces. In such cases, precise models are impractical and are too expensive. Furthermore, the relevant available information is usually in the form of empirical prior knowledge and input and output data representing instances of the systems behavior. Therefore, one needs an approximate reasoning system which is capable of handling such imperfect information.

While Bezdek defines such approaches within a frame called Computational Intelligence, Zadeh explains the same using the Soft Computing paradigm. According to Zadeh, in contrast to traditional Hard Computing, Soft Computing is tolerant of imprecision, uncertainty, and partial truth.

From the research article, the researcher expresses that large amount of data which is collected and stored for analysis. Making appropriate use of these data often leads to considerable gains in efficiency and therefore economic advantages. There are several applications of Machine Learning techniques in the field of agriculture. The researchers implemented K-Means algorithm to forecast the pollution in the atmosphere, the K Nearest Neighbor is applied for simulating daily precipitations and other weather variables and different possible changes of the weather scenarios are analyzed using Support Vector Machines. Soil profile descriptions were proposed by the researcher for

classifying soils in combination with GPS based technologies. They were applied K-Means approach for the soil classification.

In a similar approach, crop classifications using hyper spectral data was carried out by adopting one of the Machine Learning approach i.e. Support Vector Machines. One of the researchers used an intensified fuzzy cluster analysis for classifying plants, soil and residue regions of interest from GPS based color images. In the agricultural science, clustering techniques are found in grading apples before marketing. Weeds were detected on precision agriculture. The researchers worked on rainfall variability analysis and its impact on crop productivity. The effect of observed seasonal climatic conditions such as rainfall and temperature variability on crop yield prediction was considered through an empirical crop model. Furthermore, there are two approaches to investigate the impact of climate change on crop production which include the crop suitability approach and the production function approach. Machine learning in Agriculture is a Novel field still a lot of work has been done in field of Agriculture using Machine learning.

In a proposed yield prediction model which used Machine Learning techniques for classification and Prediction. This model worked on input parameters crop name, land area, soil type, soil pH, pest details, weather, water level, seed type and this model predicted the plant growth and plant diseases and thus enabled to select the best crop based on weather information and required parameters. She proposed an approach which used unsupervised learning technique K Means Clustering technique to classify the soils into clusters based on the salinity factors. This work classified the soils as Sodic, Saline -Sodic and Acidic. This model enabled the analysts to select the best soil for crop productivity She proposed a crop yield prediction model that implanted two Machine Learning techniques namely Multiple Linear Regression and Density Based Clustering techniques.

The predict answers were Year', 'Rainfall', 'Area of Sowing', 'Yield', 'Fertilizers' (Nitrogen, Phosphorous and Potassium) and Response variable was 'Production'. In Kg/Hectares. A. conducted a study on the different Machine Learning techniques used in Agriculture. The techniques like K Means.KNN, ANN, SVM were studied related to Agriculture field and concluded that these techniques in combination with GPS and Remote sensing techniques can be used to study the characteristics of soil, classify soils, classify crops and for prediction too.

III. METHODOLOGY

1) Problem Study: A brief study of problems related to maximization of the productivity and prediction of crop yield will be done by going through the related literature review, and with the brief discussions with soil analysts and farmers and broader view of research problem will be gained.

2) Data Collection: After completing the problem study and gaining insight of research problem the related data will be collected from respective government department database and respective research center. The dataset will consist of Soil Nutrient status individual field wise. data will be divided for training and testing purposes.

Data will be preprocessed and will be transformed into two excel sheets one for training and one for testing purposes.

3) Parameter Study: The Dataset collected will consists of soil composition parameters and is one of subsets for the prediction of yield. The data will consists of 12 parameters out of Sample no, Ph, EC, OC, N, P, K, S, Cu, Fe, Zn, Mn out of which 7 parameters(Ph,EC,OC,N,P,K,S) are classified as Macro-Nutrients and remaining 4 parameters (Cu,Fe,Zn,Mn)are Micro-Nutrients.

4) Training Data Categorization: The individual tuple values of each parameter will be classified into LOW,HIGH and MEDIUM category based on critical limits defined by soil chemists for soils in Vidarbha Region.

5) Applying Machine Learning Approaches:

a) Random Forest: Let assume that the user knows about the construction of single classification trees. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning

When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

b) ID3 Algorithm: In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains.

The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H(S) (or information gain IG(S)) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split or partitioned by the selected attribute

to produce subsets of the data. (For example, a node can be split into child nodes based upon the subsets of the population whose ages are less than 50, between 50 and 100, and greater than 100.) The algorithm continues to recur on each subset, considering only attributes never selected before.

Recursion on a subset may stop in one of these cases:

- Every element in the subset belongs to the same class; in which case the node is turned into a leaf node and labelled with the class of the examples.
- there are no more attributes to be selected, but the examples still do not belong to the same class. In this case, the node is made a leaf node and labelled with the most common class of the examples in the subset.
- there are no examples in the subset, which happens when no example in the parent set was found to match a specific value of the selected attribute. An example could be the absence of a person among the population with age over 100 years. Then a leaf node is created and labelled with the most common class of the examples in the parent node's set.

Throughout the algorithm, the decision tree is constructed with each non-terminal node (internal node) representing the selected attribute on which the data was split, and terminal nodes (leaf nodes) representing the class label of the final subset of this branch.

The ID3 algorithm is used by training on a data set S to produce a decision tree which is stored in memory. At runtime, this decision tree is used to classify new test cases (feature vectors) by traversing the decision tree using the features of the datum to arrive at a leaf node. The class of this terminal node is the class the test case is classified as.

IV. System Architecture

The proposed system will be designed while keeping rain statics in mind for corresponding soil. The rain data will be merged with soil data so that it will be easy to map the soil for the particular crop. The merged data will be stored separately and will be provided directly to the algorithm. The crop data will be the input for the proposed system. The crop data will be preprocessed before applying algorithm on that. The processed crop data and soil data with rain data will be processed and run through the proposed algorithm. The graphical result will be predicted which will be reported and predicted result will be displayed.

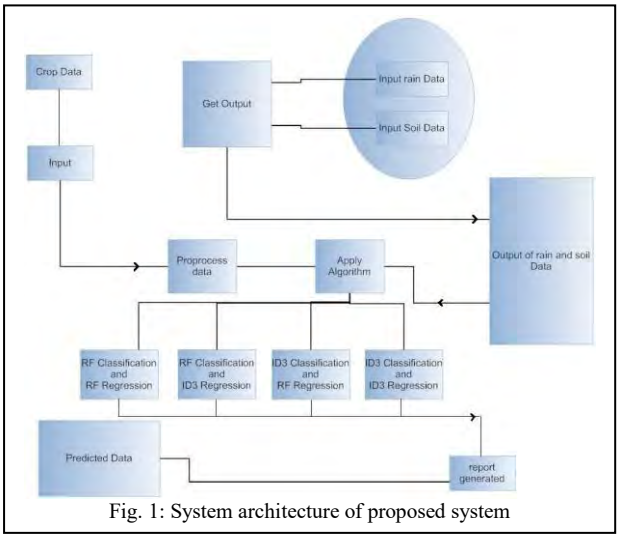


Fig. 1: System architecture of proposed system

V. Expected Outcome

The predicted result will consist of bar graph representing the success rate of predicted crop, macro nutrient and micro nutrient, rain fall history in particular soil.

VI. Conclusion

In this proposed system, by using the method of classification using ID3 combined with Random Forest Algorithm will help farmers in Vidarbha Region in deciding the crop to plough and the amount of fertilizer to be used in farm. This will surely help farmers and will enable them to be independent.

References

[1] Supriya D M "Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining Approach" in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 5, May 2017.

[2] Vaneesbeer Singh, Abid Sarwar "Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach" in International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May – June 2017.

[3] Profile of Maharashtra and selected districts, shodhganga.inflibnet.ac.in/bitstream/10603/121515/13/13_chapter4.pdf.

[4] Profile of Maharashtra and selected districts, shodhganga.inflibnet.ac.in/bitstream/10603/121515/13/13_chapter4.p df.

[5] Profile of Maharashtra and selected districts, shodhganga.inflibnet.ac.in/bitstream/10603/121515/13/13_chapter4.p df.

[6] Andrew.W "Moore Professor School of Computer Science Carnegie Mellon University", Naïve Bayes Classifiers, www.cs.cmu.edu/~awm awm@cs.cmu.edu

[7] Andrew.W "Moore Professor School of Computer Science Carnegie Mellon University", Naïve Bayes Classifiers, www.cs.cmu.edu/~awm awm@cs.cmu.edu

[8] B. Bhattacharya, D.P. Solomatine "Machine learning in soil classification" Elsevier 2006 Special Issue, Neural Networks 19 (2006) 186–195

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio "Generative Adversarial Nets"

*[10]Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems. **52**: 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377*

[11] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering" (PDF). Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284&ndash, 292. ISBN 0-521-64298-1. MR 2012999.

[12] Coates, Adam; Ng, Andrew Y. (2012). "Learning feature representations with k-means" (PDF). In G. Montavon, G. B. Orr, K.-R. Müller. *Neural Networks: Tricks of the Trade*. Springer.

[13] Csurka, Gabriella; Dance, Christopher C.; Fan, Lixin; Willamowski, Jutta; Bray, Cédric (2004). *Visual categorization with bags of keypoints* (PDF). ECCV Workshop on Statistical Learning in Computer Vision.

[14] Coates, Adam; Lee, Honglak; Ng, Andrew Y. (2011). *An analysis of single-layer networks in unsupervised feature learning* (PDF). International Conference on Artificial Intelligence and Statistics (AISTATS). Archived from the original (PDF) on 2013-05-10.

[15] Schwenker, Friedhelm; Kestler, Hans A.; Palm, Günther (2001). "Three learning phases for radial-basis-function networks". *Neural Networks*. **14** (4–5)

International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-05, Issue-04, July 2019

Comparative study on Enhanced Random Forest Algorithm by Crop Prediction

[*]Priyankar Ravindra Tiwari, [#]Dr. Anand Khandare

[*]M.E. Scholar, [#]Assistant Professor, Thakur College of Engineering and Technology, Mumbai, India, [*]priyankartiwarimy@gmail.com, [#]anand.khandare1983@gmail.com

Abstract: Earlier yield prediction was performed by considering the farmer's experience on a particular field, weather and crop. Since farmers don't have knowledge about the presence of the nutrients and they don't have the idea about the crop to plough and pesticides to be used due to which performance of agriculture is degrading in economy and farmers getting into loss which they have to pay from their own pocket, which is also a root cause of farmers suicide. This makes the problem of predicting the yielding of crops an interesting challenge.

This work presents a system, which uses machine learning techniques called Random Forest with enhanced performance in order to predict the category of the analysed soil datasets. The category, thus predicted indicates the yielding of crops. The problem of predicting the crop yield is formalized as a classification and regression rule, where Enhanced Random Forest is divided into classification and regression, used for categorization of soil and predicting rainfall and both the result will provide the predicted crop to plough.

Keywords — Random Forest, classification, regression, enhanced Random Forest, decision tree, ID3, crop prediction

I. INTRODUCTION

Agriculture is the backbone of the Indian. The agriculture data increases day by day. Since an outsized population lives in rural areas and is directly or indirectly captivated with agriculture for a living. Outlay from farming forms the main source for the farming community. The essential requirements for harvesting are water resources and ability to buy seeds, fertilizers, pesticides, labour etc. Most farmers raise the required capital by compromising on other essential expenditures, and when it is still insufficient, they resort to credit from sources like banks and private commercial institutions. In such a situation, the repayment is dependent on the success of the harvest. If the harvest fails even once because of many factors, like atmospheric condition pattern; soil type; improper, excessive, and ill-timed application of each fertilizers and pesticides; debased seeds and pesticides etc. Most power of soil in nature comes from soil survey efforts. Soil survey, or soil mapping, is the process of determining available nutrients in soil or other holding of the soil cover over a landscape, and mapping them for others to understand and use. Primary data for the soil survey is acquired by area sampling and supported by remote sensing. As the volume of data increase, it requires involuntary way for these data to be extracted when needed. Machine Learning can be used for pretend the next trends of agricultural processes. Every soil is a mixture of these component: Nitrogen, Phosphorus, Potassium, pH Value and Electrical Conductivity. Based on these factors we predict the soil fertility level and crop for a particular soil sample.

In this context, the goal of this paper is to provide a comprehensive, comparative and self-contained analysis of a class of algorithms known as ID3 decision trees and random forests and enhanced random forest. These methods have proven to be a robust, accurate and successful tool for solving countless of machine learning tasks, including classification, regression, density estimation, manifold learning or semi-supervised learning.

1.1 Problem Definition:

A brief study of problems related to maximization of the productivity and prediction of crop yield has been done by going through the related literature review, and with the brief discussions with soil analysts and farmers and broader view of research problem has been gained. Yield prediction is incredibly well-liked among farmers currently, that notably contributes to the right choice of crops for sowing. This makes the problem of predicting the yielding of crops an interesting challenge. Earlier yield prediction was performed by considering the farmer's expertise on a selected field and crop. This work presents a system, which uses Machine Learning techniques in order to predict the category of the analysed soil datasets. The category, thus predicted indicates the yielding of crops.

 DOI : 10.35291/2454-9150.2019.0291

1.2 Motivation:

Farmers in India, specially Vidarbha region in Maharashtra state faces drought due to which their crop and yielding is getting degraded. They don't have any idea about availability of nutrient in their field. They use their own experience to plough the crop which have very less success ratio. Due to less success ratio they are unable to pay their loan amount sanctioned for their crop. In unsuccessful for their repayment of the loan amount they attempt to suicide which is a main reason for highly rising ratio in farmers suicide.

To help the farmers to decide the crop to be plough for their benefits I am motivated to build this system. This system collects the data from the soil testing laboratory supported by Department of Agriculture, Government of India. This dataset consists of the available nutrient for farmers' soil and rainfall for particular region.

II. LITERATURE REVIEW

Random Forest (RF) is an ensemble classifier proposed by Breiman (2001) which consists of many sub-models. The predictions and other quantities of interest are obtained by combining the outputs of all the sub-models. The sub-models for Random Forest are classification and regression trees (CART) which is the key for understanding the Random Forest.

In the past decade, various methods have been proposed to grow a random forest (Breiman, 2001; Dietterich, 2000; Ho, 1998). Among these methods, Breiman's method (Breiman, 2001) has gained increasing popularity because it has higher performance against other methods (Banfield et al., 2007).

Let D be a training dataset in an M-dimensional space X, and let Y be the class feature with total number of c distinct classes. The method for building a random forest (Breiman, 2001) follows the process including three steps (Baoxun Xu et al., 2012):

Step 1: Training data sampling: use the bagging method to generate K subsets of training data {D1, D2, ..., DK} by randomly sampling D with replacement;

Step 2: Feature subspace sampling and tree classifier building: for each training dataset Di ($1 \leq i \leq K$), use a decision tree algorithm to grow a tree. At each node, randomly sample a subspace Xi of F features ($F << M$), compute all splits in subspace Xi, and select the best split as the splitting feature to generate a child node. Repeat this process until the stopping criteria is met, and a tree hi(Di, Xi) built by training data Di under subspace Xi is thus obtained;

Step 3: Decision aggregation: ensemble the K trees {h1(D1, X1), h2(D2, X2),...hK(DK, XK)} to form a random forest and use the majority vote of these trees to make an ensemble classification decision. (i.e., majority votes for classification, average for regression).

The algorithm has two key parameters, i.e., the number of K trees to form a random forest and the number of F randomly sampled features for building a decision tree. According to Breiman (2001), parameter K is set to 100 and parameter F is computed by F=[log2M + 1]. For large and high dimensional data, a large K and F should be used.

The estimation of the error rate can be obtained based on the training data as follows:

1. At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls "out-of-bag", or OOB data) using the tree grown with the bootstrap sample.
2. Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate.

2.1 Advantages of Random Forest:

1. Accuracy is as good as Adaboost and sometimes better.
2. It is faster than bagging or boosting.
3. It gives useful internal estimates of error, strength, correlation and variable importance.
4. It is simple and easily parallelized

2.2 Disadvantage of Random Forest:

1. Models in Random Forest which has been overfit will have poor predictive performance as it doesn't generalize well. Generalization means how well model makes prediction for the cases that are not in training set.
2. In Random Forest Algorithm we need to choose number of trees.
3. Large number of attributes for prediction and large number of trees makes algorithm slower.
4. For data including categorical variables with different number of levels, random forests are biased in favour of those attributes with more levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

2.3 Related Work:

Over the past decade, some research was invested in boosting the performance of RF. One of the earliest to be reported is by Latinne *et al.* (2001). A method based on the McNemar non-parametric test of significance was proposed. The method a priori determines the minimum number of trees in the RF to use in order to obtain prediction accuracy comparable to the one obtained with larger ensembles. In addition to maintaining accuracy with fewer trees, the method significantly improves classification speed and reduces memory costs.

Robnik-Šikonja (2004) investigated new ways to improve the performance of RF. By using several attribute evaluation measures instead of just one, the correlation

 DOI : 10.35291/2454-9150.2019.0291

between trees is decreased without any loss in their strength. Another way to improve the performance of RF is to change the voting method. Instead of using majority voting, weighted voting is used. With this voting technique, internal estimates are used to identify instances most similar to the instance being labeled. The votes of the corresponding trees are then weighted with the strength they demonstrate on these near instances. Improvements were demonstrated on several classification data sets.

Tsymbal *et al.*, (2006) found a way to improve the performance of RF on some data sets by replacing majority voting with more sophisticated dynamic integration techniques. Three techniques were used: Dynamic Selection (DS), Dynamic Voting (DV), and Dynamic Voting with Selection (DVS). Using DV and DVS integration strategies, experimental studies showed that dynamic integration was able to improve the accuracy of RFs on 12 out of 27 data sets.

III. METHODOLOGY

Improving accuracy in classification and prediction has been grasping a lot of attention from many researchers all over the world. Random Forest is a new approach to data exploration, data analysis, and predictive modelling. This research work focuses on improving the performance of random forest in three aspects.

3.1 Enhanced Random Forest Algorithm:

The standard algorithm has two key parameters, *i.e.*, the number of *n* trees to form a random forest and the number of *F* randomly sampled features for building a decision tree. According to Breiman (2001), parameter *K* is set to 100 and parameter *F* is computed by F=[log2M + 1].

To enhance the algorithm, the samples feature should not be limited. For this, in enhanced algorithm number of F max_feature is randomly is entered by algorithm and the best one is selected for the system. Also the standard algorithm need to be entered the number of "*n*" Trees to a random forest and the larger the number of the trees slower the speed of algorithm. Thus, to resolve this enhanced algorithm is implemented with random function which randomly choose the number of trees and checks the result for each and select the number of tree which has the best result and uses that for all the further steps.

1. Randomly select "k" features from total "m" features.
   Where k << m
2. Among the "k" features, check for every feature and select "f" best feature.
3. For the feature "f" calculate the node "d"
4. Split the node "d" into daughter node "l" using best split.
5. Repeat 1 to 3 steps until "l" number of nodes has been reached.
6. Randomly put the value of number of trees and select "n" the number giving best result
7. Build forest by creating "n" number of trees.

The above algorithm generates forest by following above algorithm which is enhanced to improve performance. After building forest, classification and regression is applied as below:

1. Draw n_{tree} bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when, m_{try} = p, the number of predictors.)
3. Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

3.2 Flowchart of Enhanced Random Forest Algorithm:

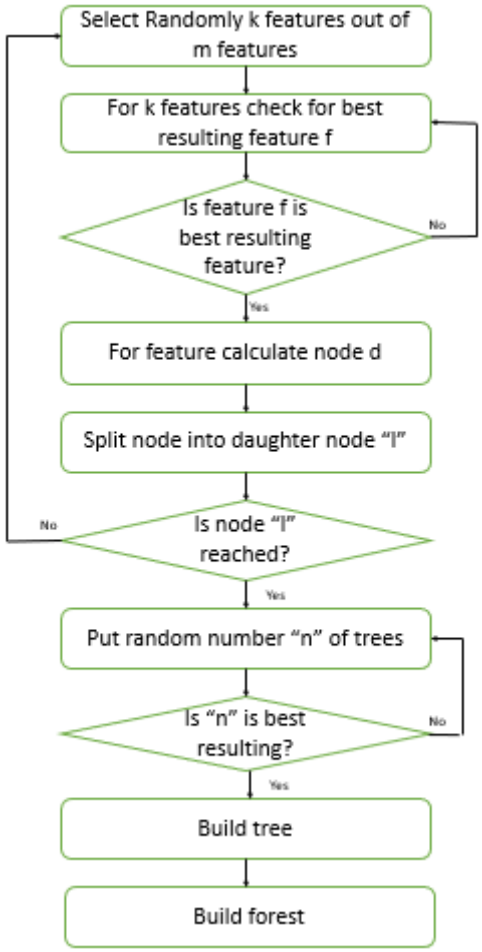


Fig 1: Flowchart of Enhanced Random Forest Algorithm

IV. RESULT

The algorithm measure performance on different parameters. Few parameters like accuracy, OOB Error Rate, Confusion Matrix, Error Rate, Mean Squared Error, R2 Score.

 DOI : 10.35291/2454-9150.2019.0291

4.1 Dataset:

Datasets used in this system numeric value which consist of the nutrient value in different unit and rainfall in different area is recorded in mm (Millimeter). The soil dataset consist of the nutrient value for the classification of fertility level. There are total 880 soil sample have been gathered and based on that these data sets have been prepared. The dataset have been devided into 80:20 for training and testing respectively.

The rainfall dataset consist of rainfall record of maharashtra region in different state. There are 3168 record have been used for the prediction and it also have been devided into 80:20 for training and testing. The rainfall data contains rainfall record month wise from year 2010 to 2017. It is used to predict the rainfall for the required year and month.

Crop dataset which records the list of crop, required minimum and maximum rainfall and fertility level of the soil in which it can be grown.

4.2 Result for Standard Random Forest:

```
D:\ME PROJ\Final>py rf_rf.py
Enter value of N : 333
Enter value of P : 7.5
Enter value of K : 507
Enter value of ph : 7.53
Enter value of ec : 0.54
Enter value of Disctrict : Pune
Enter value of year : 2019
Enter value of month : July
Train set
Random Forest:Confusion Matrix:
 [[317  11   0]
 [ 10 339   0]
 [  1  26   0]]
Random Forest OOB error rate : 0.9147727272727273
AUC for random forest: 0.8007215747072657
Accuracy For Random Forest: 0.9318181818181818
Error rate for random forest: 0.06818181818181823
Test set
Random Forest:Confusion Matrix:
 [[80  4  0]
 [ 3 79  0]
 [ 0 10  0]]
Random Forest OOB error rate : 0.9147727272727273
AUC for random forest: 0.789041786777059
Accuracy For Random Forest: 0.9034090909090909
Error rate for random forest: 0.09659090909090906
Train set
Random Forests Mean squared Error :37863.22479915654
Random Forests r2_score :0.5223538960378842
Random Forest Error Rate :0.4776461039621158
Tesst set
Random Forests Mean Squared error :19369.758205601724
Random Forests r2_score :0.4915544808788034
Random Forest Error Rate :0.5084455191211966
```

Fig 2: Standard Random Forest Result

Table 1: Standard Random Forest Performance

Random Forest Classification						
ROC			OOB Error Rate	AUC	Accuraacy	Error Rate
Class 0	Class 1	Class 2				
0.99	0.99	0.89	0.91	80.07	93.18	0.07
Random Forest Regression						
Mean Squared Error			**R^2 Score**			**Error Rate**
19369.75			**0.49**			**0.51**

The predicted values also shown in above snap. Predicted crop will be displayed in another window as shown below.

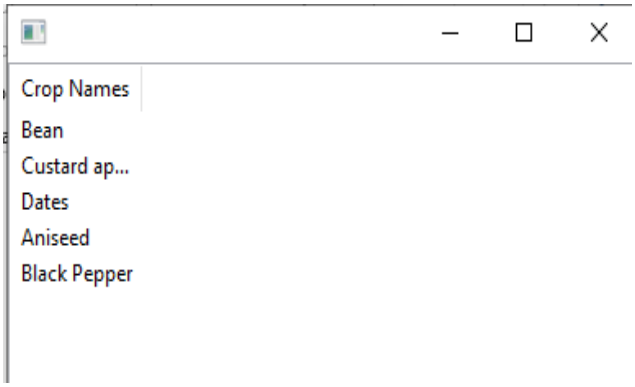


Figure 3: Predicted Crop by Standard RF

4.3 Result for Enhanced Random Forest:

```
Enter value of N : 333
Enter value of P : 7.5
Enter value of K : 507
Enter value of ph : 7.53
Enter value of ec : 0.54
Enter value of Disctrict : Pune
Enter value of year : 2019
Enter value of month : July
Train set
Random Forest:Confusion Matrix:
 [[316  11   0]
 [  7 337   0]
 [  1  30   2]]
Random Forest OOB error rate : 0.9090909090909091
AUC for random forest: 0.8119181719376315
Accuracy For Random Forest: 0.9303977272727273
Error rate for random forest: 0.06960227272727271
Test set
Random Forest:Confusion Matrix:
 [[79  6  0]
 [ 2 85  0]
 [ 0  4  0]]
Random Forest OOB error rate : 0.9090909090909091
AUC for random forest: 0.795347614403156
Accuracy For Random Forest: 0.9318181818181818
Error rate for random forest: 0.06818181818181823
Train set
Random Forests Mean squared Error :20170.064626777286
Random Forests r2_score :0.7520845463421723
Random Forest Error Rate :0.24791545365782774
Tesst set
Random Forests Mean Squared error :9336.364855780997
Random Forests r2_score :0.6832571953329671
Random Forest Error Rate :0.31674280466703286
Predicted Values Using Random Forest as Classification and Random Forest Regression :

Predicted Fertility using random forest classification :
 [1]
Predicited rainfall using random forest regression :
 [99.95907572]

D:\ME PROJ\Final>
```

Figure 4: Performance Measurement for Enhanced RF

Table 2: Performance table for Enhanced Random Forest

Random Forest Classification						
ROC			OOB Error Rate	AUC	Accuraacy	Error Rate
Class 0	Class 1	Class 2				
1.00	0.99	0.98	0.90	79.54	93.18	0.07
Random Forest Regression						
Mean Squared Error			**R^2 Score**			**Error Rate**
9336.36			**0.68**			**0.31**

 DOI : 10.35291/2454-9150.2019.0291

Figure 5: Predicted crop using Enhanced RF

4.4 Comparison:

Table 3: Comparison of RF and ID3 against Enhanced RF

		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
Enhanced Random Forest	Error Rate	0.25	0.21	0.23	0.23	0.21	0.22	0.2	0.24	0.21	0.23
ID3		0.52	0.65	0.49	0.49	0.54	0.51	0.57	0.51	0.49	0.52
Standard Random Forest		0.51	0.48	0.5	0.52	0.57	0.47	0.49	0.54	0.51	0.48
Enhanced Random Forest	R2 Score	68.32	70.42	71.64	72.48	65.56	78.56	78.48	68.45	67.58	72.65
ID3		50.26	48.46	49.64	48.65	47.35	49.85	47.67	51.65	52.26	48.65
Standard Random Forest		49.15	50.12	50.1	48.16	50.16	50.75	49.26	46.82	50.1	51.46
Enhanced Random Forest	Mean Squared Error	9336.36	7645.56	7546.3	8343.16	9435.23	7653.64	8323.8	8542.35	9356.64	7564.65
ID3		19475.35	20365.95	18356.24	17356.32	19648.65	20315.94	17546.65	18356.26	17365.32	18653.23
Standard Random Forest		19369.76	15647.23	17534.35	16547.24	13247.23	16326.23	17453.12	17354.32	18975.64	19346.64
Enhanced Random Forest	Error Rate	0.07	0.08	0.08	0.07	0.08	0.08	0.08	0.08	0.07	0.08
ID3		0.08	0.09	0.09	0.08	0.09	0.1	0.09	0.09	0.1	0.09
Standard Random Forest		0.09	0.09	0.08	0.08	0.09	0.09	0.1	0.09	0.08	0.08
Enhanced Random Forest	Accuracy	93.2	93.4	92.3	92.8	92.2	93.4	92.1	92.2	92.4	92.8
ID3		89.3	90.2	91.5	91.2	89.5	89.6	88.7	90.1	90.2	89.2
Standard Random Forest		90.3	91.2	91.7	91.9	90.7	90.6	91.2	90.5	89.3	91.8
Enhanced Random Forest	AUC	0.81	0.8	1	0.9	0.97	0.98	0.9	0.91	0.97	0.94
ID3		0.72	0.76	0.8	0.93	0.94	0.72	0.94	0.9	0.81	0.85
Standard Random Forest		0.79	0.78	0.9	0.99	0.95	0.79	1	0.89	0.88	0.86
Enhanced Random Forest	OOB Error rate	0.9	0.9	0.89	0.9	0.88	0.87	0.88	0.85	0.87	0.88
ID3		0.91	0.9	0.9	0.91	0.92	0.89	0.9	0.88	0.91	0.9
Standard Random Forest		0.91	0.89	0.92	0.88	0.92	0.9	0.87	0.92	0.9	0.89

V. DISCUSSION

As compared in comparison table, Enhanced Random forest classification algorithm gives higher ROC Value, Less OOB Error Rate, higher AUC, Accuracy and less Error Rate against Random Forest Regression in comparison to ID3 algorithm as regression. Even ID3 didn't perform well neither against Random Forest nor Enhanced Random Forest algorithm. Also, in Comparison table, Enhanced Random Forest perform very well. It returns adequate lower Mean Squared Error compared to ID3 regression against Random Forest algorithm. Also, it returns with the higher R^2 Score and lower Error rate in compared to any other classification algorithm.

VI. FUTURE SCOPE

The future of the Random Forest as classification and Regression involves predicting the pesticides and fertilisers to be used to improve fertility level of soil based on current micro and macro nutrient available in soil. Random Forest as classification and Regression is also helpful in predicting the rainfall for coming years based on previous rainfall trend.

VII. CONCLUSION

Earlier yield production was decided based on farmers experience where technology involvement was not there which gives accurate answer to decide the crop to plough. Therefore, in order to help farmers to decide the crop to plough for their financial as well as social benefits crop prediction system make use of Random Forest as classification as well as regression. Classification algorithm classifies the soil sample based on the available nutrient in soil into different class of soil where as regression predicts the expected rainfall for the entered year and month in which farmer want to plough.

Enhanced Random Forest classification and regression which performed in comparison. The classification comparison is based on the parameter ROC Curve, AUC, OOB Error Rate, Accuracy and Error Rate, where as Regression comparison is based on parameter Mean Squared error, R^2 Score and Error rate.

The planned model work presents comparison of Random forest Classification combined with Random Forest Regression and ID3 Regression and Enhanced Random Forest. Different soil sample have been used to compare the algorithm and it is concluded that Enhanced Random Forest as classification and Regression performed better in term of RUC Curve, AUC, Accuracy, Error Rate and OOB Error Rate. Accuracy is most important parameter that demonstrate the performance of any algorithm. It is observed that accuracy of Enhanced Random Forest as Classification and Regression combined is better in classifying the dataset and predicting the result.

 DOI : 10.35291/2454-9150.2019.0291

REFERENCES

[1] Supriya D M "Analysis Of Soil Behavior And Prediction Of Crop Yield Using Data Mining Approach" In International Journal Of Innovative Research In Computer And Communication Engineering, Vol. 5, Issue 5, May 2017.

[2] Vaneesbeer Singh, Abid Sarwar "Analysis Of Soil And Prediction Of Crop Yield (Rice) Using Machine Learning Approach" In International Journal Of Advanced Research In Computer Science, Volume 8, No. 5, May – June 2017.

[3] Profile Of Maharashtra And Selected Districts, Shodhganga.Inflibnet.Ac.In/Bitstream/10603/121515/13/13_Chapter4.Pdf.

[4] Andrew.W "Moore Professor School Of Computer Science Carnegie Mellon University", Naïve Bayes Classifiers, Www.Cs.Cmu.Edu/~Awm Awm@Cs.Cmu.Edu.

[5] Andrew.W "Moore Professor School Of Computer Science Carnegie Mellon University", Naïve Bayes Classifiers, Www.Cs.Cmu.Edu/~Awm Awm@Cs.Cmu.Edu.

[6] B. Bhattacharya, D.P. Solomatine "Machine Learning In Soil Classification" Elsevier 2006 Special Issue, Neural Networks 19 (2006) 186–195.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio "Generative Adversarial Nets"

[8] Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (Black) Art Of Runtime Evaluation: Are We Comparing Algorithms Or Implementations?". Knowledge And Information Systems. 52: 341–378. Doi:10.1007/S10115-016-1004-2. Issn 0219-1377

[9] Mackay, David (2003). "Chapter 20. An Example Inference Task: Clustering" (Pdf). Information Theory, Inference And Learning Algorithms. Cambridge University Press. Pp. 284&Ndash, 292. Isbn 0-521-64298-1. Mr 2012999.

[10] Coates, Adam; Ng, Andrew Y. (2012). "Learning Feature Representations With K-Means" (Pdf). In G. Montavon, G. B. Orr, K.-R. Müller. Neural Networks: Tricks Of The Trade. Springer.

[11] Csurka, Gabriella; Dance, Christopher C.; Fan, Lixin; Willamowski, Jutta; Bray, Cédric (2004). Visual Categorization With Bags Of Keypoints (Pdf). Eccv Workshop On Statistical Learning In Computer Vision.

[12] Coates, Adam; Lee, Honglak; Ng, Andrew Y. (2011). An Analysis Of Single-Layer Networks In Unsupervised Feature Learning (Pdf). International Conference On Artificial Intelligence And Statistics (Aistats). Archived From The Original (Pdf) On 2013-05-10.

[13] Schwenker, Friedhelm; Kestler, Hans A.; Palm, Günther (2001). "Three Learning Phases For Radial-Basis-Function Networks". Neural Networks. 14 (4–5): 439–
458. Citeseerx 10.1.1.109.312. Doi:10.1016/S0893-6080(01)00027-2

[14] Geetha Mcs. Implementation Of Association Rule Mining For Different Soil Types In Agriculture. International Journal Of Advanced Research In Computer And Communication Engineering. 2015 Apr; 4(4):520–2.

[15] Knowledge Discovery And Data Mining To Identify Agricultural Patterns, Kulwant Kaur, Maninderpal Singh, Ijesrt [1337-1345], March, 2014

[16] G.Kesavaraj, Dr.S.Sukumaran "A Study On Classification Techniques In Data Mining" Ieee – 31661.

 DOI : 10.35291/2454-9150.2019.0291

Acknowledgement

I would like to take the opportunity to express my sincere thanks to my **Dr. Anand Khandare** Assistant Professor, CMPN Department, TCET for his invaluable support and guidance throughout my P.G. research work. Without his kind guidance & support this was not possible. I am grateful to him for timely feedback which helped me track and schedule the process effectively. His time, ideas and encouragement that he gave helped me to complete my project efficiently.

I would like to thank **Dr. R.R. Sedamkar** for his guidance and encouragement throughout my Post Graduation.

I would also like to thank **Dr. B. K. Mishra**, Principal, Thakur College of Engineering and Technology, for his encouragement and for providing an outstanding academic environment, also for providing the adequate facilities. I am thankful to all my M.E. teachers for providing advice and valuable guidance.

I also extend my sincere thanks to all the faculty members and the non-teaching staff and friends for their cooperation. Last but not the least, I am thankful to all my family members whose constant support and encouragement in every aspect helped me to complete my project.

Priyankar Ravindra Tiwari