International Journal of Modern Agriculture, Volume 10, No.2, 2021
ISSN: 2305-7246

A Random Forest Algorithm for Predicting Crop Yield in Hilly Regions of North East India

Bimal Deb Nath[1] and Lipika Sarkar[2]
[1]Assistant Professor, Dept. of Management, NEHU, Tura Campus, Meghalaya, India
[2]Research Assistant, ICSSR project under Impress 2019
E-Mail: bimal.dn@nehu.ac.in, lipika.etx@gmail.com

Abstract:
The North-Eastern Region (NER) of India, despite the potentialities of valuable natural resources, is struggling in the agricultural sector. The sector is cropped up with multiple problems of decline in agricultural products such as soil problems, land and water degradation, low potential areas of eastern India, marketing and finance issues. The region could not nourish its resources due to diverse natural location, lack of proper attention, and significant yield prediction. However, crop yield prediction being one of the critical factors in agriculture practices, there is a need to generate relevant area-specific information for high yield. The farmers in the region are unaware of the challenges and opportunities in agriculture due to a lack of adequate information, and therefore, these farmers largely need information regarding crop yield before sowing seeds to achieve enhanced crop yield. Given the advantages offered from accurate predictions of crop yields and the ability of data mining techniques to extract patterns in large data sets, this paper initially focuses on major influencing factors of crop production and then determines how accurately the random forest, i.e., a data mining technique, can estimate crop productivity in the hilly region of North East India in the pursuit of advancing industry sustainability. According to the experiment on testing data sets, the method has high classification accuracy and it is more suitable for the current big data scene in which data patterns will gradually change with time.
Keywords: Agriculture, Crop Yield Prediction, Data Mining, Random Forest, North East India.

Introduction:

In many developing countries, with a particularly rapidly growing population, agriculture plays a vital role in the economy and the stable income generation (Lee, 2005; Birthal et.al. 2006; Kannan et.al. 2011). In these countries, agriculture is the backbone of the economic system and is the core basis for livelihood and poverty alleviation. It also provides employment opportunities to a substantial percentage of the population (Deb, 1994) and has a vital role in sustainable development and GDP contribution. Thus, the development of the agricultural sector in these counties, including India, has been crucial (Aker, 2011). Statistics reveal that India lies at fourth position in the list of leading agricultural countries across the globe. Besides, it is considered that around 85.4% of the Indian population is associated with the agriculture sector for their livelihood (Anantharaman et.al. 2016). In India, largely, North–Eastern Region (NER) is favorable for the record of crop diversification and the region is home to few of the high-value niche crops like joha rice, Assam lemon, medicinal rice, sticky rice (Manipur), pineapple, aromatic rice of Tripura, khasi mandarin, ginger, turmeric (Meghalaya), french bean, king chili, bird eye chili (Nagaland), Assam tea, large cardamom, ginger, pineapple and passion fruits and these crops. The other strengths of NER, India, are that due to thin population density compared to vast natural resources available in the region, there is less pressure on food demand.

However, the NER, India, despite the potentialities of valuable natural resources compared to the rest of the country, is struggling in the agricultural sector (Banerjee, 2006), and the indigenous farming systems are still prevailing in the region (Das et al. 2018). With the increase in population pressure, there is a need for

technological backstopping or mass adoption of modern techniques and advancements for sustainable crop production (ICAR, 2013). Adopting intelligent agriculture driven by information technology and good management practices, and efficient decision-making is crucial to boost agricultural yield growth and limit environmental externalities in the region (ICAR, 2013; Vetter et.al. 2017; Mythili&Goedecke, 2016).

In agriculture, actual time and historically generated data are generated by on-site farming, satellite farming, etc. in an unstructured and structured data format called precision agriculture (Bharadi et.al. 2017). Research affinity is currently to find out the knowledgeable information of the data gathered from the different sources of precision agriculture for predictive decisions (Manjula et.al. 2016; Shettar & Angadi, 2016). The agricultural sector required enhancement in the process of decision-making that can be used to increase the quantity of data and information, which comes from a broader number of different resources (Bharadi et.al. 2017; Manjula et.al. 2016). With the increase in technology, innovative devices /equipment are used in farming, and everything is digitized, which leads the agricultural data to enter into the world of big data (Grisso et.al. 2009; Bharadi et.al. 2017) and with all this information, farmers can avoid significant loss and subsequently can be benefited applying the information effectively (Bharadi et.al. 2017; Narkhede & Adhiya, 2014).

The data mining approach for processing huge data such as big data could help the agricultural firms to increase one of their production practices, such as acquiring new farmers by framing new profitable agriculture schemes based on crop yield prediction and eventually campaigning to different groups of farmers about a particular scheme based on the result of the grouping of various crops depending on standard features (Sudha et.al. 2018). In big data scenarios, classifier technology is one focus of data mining research as the complexity of classification rapidly increasing, thus rejecting the applicability of the traditional classification algorithm such as association rules (Liu et.al. 2001), bayes (Bernardo & Smith, 2001), decision trees (Liu et.al. 2002), neural networks, K-means, genetic algorithms, rough sets, fuzzy logic (Li et.al. 2010; Li, 2008). As a standard method of data mining, the random forest (R.F) method (Breiman, 2001) has been proved to be a state-of-the-art learning model, which has a good classification and regression performance and fast and efficient operations. The random forest algorithm can effectively handle multiple classification problems, also has an obvious advantage in dealing with noise. The random forest method that is not subject to memory limitations and featured rapid processing speed and good parallel scalability is an excellent classification tool to handle massive data and a typical decision tree classification algorithm. Therefore, a random forest (R.F.) algorithm can be helpful for crop prediction based on the previous crop sequences in the same farmland with the current soil nutrient information. However, in the real-life world, in a problem like agriculture, R.F. faces many challenges, like selecting different parameters for crop parameters. The present study attempts to address the prediction of crop yield in NER, India to achieve preliminary information using R.F. The paper focuses on choosing the most suitable parameters of crop production and proposes a prediction model using random forest, a data mining technique to predict crop yield and its productivity in hilly regions of North East India using the knowledge of the experts.

The paper is organized as follows: initially, in section one this introduction and then section two briefly highlights related work on suitable parameters of crop production and its prediction model. Section three is devoted to research methods that include collecting data with pre-processing and building a random forest prediction model. Finally, section four argued the empirical findings with conclusions and suggestions.

Related work

Initially, some of the earlier studies relevant to the factors influencing the production of crop yields and its status concerning North-East India are highlighted, and then the related work on the crop prediction model has been reviewed thematically.

Factors influencing for Production of Crops in North East India

The crop to be produced is depended on the type of physical (including land, soil quality, topography, climate, water, location, distance, etc.) and biological elements (include crops physiology, diseases) (Arun & Sharma, 2006). As a result of atmospheric variation, crop production has taken an uneven path (Dhivya, et.al. 2017). The significant factors that may affect agricultural crop productivity in North East India were discussed below:

Crop- Soil

In India, soil characteristics are an important physical factor affecting agriculture crop productivity and differ in physical and chemical composition (Patangray et.al. 2016). The maximization of crop yield depends on the following parameters of soil such as type of soil, soil composition, and soil nature that comprises of moisture, pH of the soil, soil tilth, soil texture, soil density, soil temperature, soil color, soil depth, soil nutrient, soil consistency, soil fertility (Patel et.al. 2008; Kumar et.al. 2015). Besides, out of various factors, the soil types, nutrient, and pH were some of the critical aspects of soil fertility, being the primary source of macronutrients (N, P, K, Ca, S, Mg, C, O, H) and micronutrients (Fe, B, Cl, Mn, Zn, Cu, Mo, Ni) (Prodhan et.al. 2018). Furthermore, the crop grows very high or moderate if planted in soils that satisfy their pH levels between 4.5 and 7.5 and on the other hand, low soil pH level and lacks soil of nutrients, causes deficiencies in crops, that negatively impact crop growth and therefore decreases yield (Patangray et.al. 2016). However, soil types found in hilly regions of NER, India are primarily sandy clay loam (54% of the area), clay loam, and sandy clay, with pH, ranges from 4.41 to 5.91, and red loam and lateritic soils with pH ranges from 4.5– 6. These indicate that the pH levels of the region's soil are favorable and help crops to grow at a faster rate (Anantharaman et.al. 2016).

Crop topography

The crop-topographical factors are land preparation, land situations/suitability, land area usage, method of sowing, harvesting, and varieties (Foodstart, 2014; Liliane & Charles, 2020). One of the leading causes for low crop yield is the non-determination of land for cultivation. In hilly areas of the north-eastern region, traditional approaches of land determination are still used for better field or land preparation (Das et.al. 2016). These approaches are cutting, clearing, and burning of jungles followed by sowing, digging, and applying minimum tillage or no-tillage or dibbling methods.

Socio-Economic

The majority of farmers in NER, India, are landowners and mostly has alternative cropping. These farmers are not willing to continue farming operations in an area if they perceive that the operations will not remain economically viable (Zollinger & Krannich, 2002). The dominant farming system practiced in the region is agronomy and animal husbandry, and the intention or attitude of these farmers are sometimes not optimistic towards other crops (Ali & Das, 2017).

Crop –Climate

The environmental factors that influence the extent of crop agriculture are variations in altitude, soil, annual rainfall, weather, average temperature, the global increase of atmospheric CO2, and fluctuations in sea levels (Agrawal & Mehta, 2007; Raza et.al. 2019). The variation in these factors may lead to considerable losses of crop

production and eventually impact crop yields (Jawad et.al. 2016; Singh, 2015). One of the key environmental factors is crop-climate (Singh, 2015) and it plays a dominating role in agriculture. Crop climate interaction is one of the significant key determinants of crop yield, and the critical climatic factors that influence the growth of crops are temperature, radiation, and rainfall (Aggarwal, 2009). In India including NER, mostly climate is favorable and the crop yield production is well correlated with the monsoon rainfall and climate variability (Kumar, 1998; Neenu et.al. 2013; Roy et.al. 2014a).

Economic Factors

The crop-agricultural economic factors depend on traditional farming systems, traditional cropping system, co-operative farming, marketing area, regulation and competition, agricultural labor, and selecting the appropriate crop (Deshmukh, et.al. 2017). Agriculture, in modern times, is becoming mechanized (Pingali, 2010) with advanced machinery, fertilizers, pesticides, and high-yielding variety seeds but that require huge capital investments. However, NER India, the farmers, cannot afford to use modern farm technology due to huge capital investments. The region is facing the biggest hurdle for increasing economic activity due to lack of proper development of the secondary and tertiary sector, the infrastructural issue, lack of processing, packing, storing, and transport facilities. The efficient means of transportation widens the market for agricultural products. Therefore, once these facilities develop, the region can look forward to a prosperous future (Roy et.al. 2014b; Dabral, 2002).

Technological factors

The technological advancement influences the agricultural production of crops by overcoming constraints and introducing new developments in agriculture (Singh, 2015) and a wide range of technological innovations in agriculture like genetic improvement of varieties, fertilizer technology, pesticides, proper irrigation method, farm machinery, agronomic and management practices (integrated management of nutrients and pests) helps to encompass the crop management and cropping pattern (Kumaraswamy & Shetty, 2016; Pingali et.al. 2019). In India, including NER, India's technological advancement has replaced the previous dependence of one crop farming to multiple cropping patterns, mixed cropping, and intercropping. However, these factors pose a significant risk to farms when these are not correctly monitored or well managed (Das et.al. 2016).

Irrigation Method

The factors of crop-irrigation are quality of water, water availability, water sources, irrigation methods, drainage system duration, irrigation management, and system. To facilitate suitable moisture environment to the crops, to increase to the fertility of the soil and high crop yielding, the availability of these resources is essential to obtain optimum and sustained crop yields (Dhas et.al. 2006; Bhanja et.al. 2017). The irrigation practices permit better utilization of all production factors, leading to increased yield per unit of land. In India, farmers in the northeast mostly practice bamboo drip irrigation and continuous flow irrigation as the region gets the highest rainfall (Borthakur, 2002). However, sometimes, in the dry season, most of the farmers in the villages are depending on the existing river, a perennial stream, mountain streams, seasonal stream, open well, deep tube well, and supply of PHE water for their daily uses, cultivation practices, and rearing of livestock (Das et.al. 2016). Therefore, a proper irrigation practice must be efficiently used by the farmers for assured crop production and a sustainable agricultural system (Raman et.al. 2015).

Crop-pest

In hilly areas of NER, India, the heavy and continuous rainfall with higher humidity is congenial for increasing crop pests and diseases (Deka et.al. 2010), and the farmers must know the crop-damaging biotic agents such as weed and pests and the method to control them (Aggarwal, 2000). In this case, crop pest control methods such as natural, applied conventional, integrated pest management, local indigenous knowledge, seed treatment, weeding frequency, proper green manure, and fertilizers (Hazarika et.al. 2006) are the best way to reduce pests and diseases. However, the traditional agricultural system of NER is by, and large organic by default and the consumption of inorganic fertilizers and pesticides is very low, particularly in Sikkim, Meghalaya, Nagaland, and Arunachal Pradesh (Rahman et.al. 2007; ISPS, 2005).

Crop sustainability

To increase crop productivity and sustainability, it is very important to assess the soil fertility through the use of proper agricultural systems such as tillage, use of recommended rates, farmyard manure, nutrients, and/or crop residues into the soil and avoid sewage sludge irrigation (Wang, 2014; Singh, 2002). An appropriate application improves soil's physical properties in the long term and ensures sustainable agriculture (Shang et.al. 2014). Alternatively, agricultural practices such as organic production of crops are promoted as environmentally friendly, reducing agrarian impacts on soil and water quality (Liliane & Charles, 2020). Further, the development of integrated crop-soil system management and integrated pests management with existing crop varieties and the increase of improved and adapted high-yielding varieties under quality water and nutrient-limited environment are needed to achieve specified crop sustainability (Shah& Wu, 2019).

Given the above, the factors that affect crop yield in agriculture can be categorized into Crop-Topography, Crop-Climate, Crop-Soil Suitability, Crop-Irrigation, Crop-Pest Control, Crop-Agricultural Economics, and Crop-Sustainability Strategy.

Crop Prediction model

Predicting a crop well in advance requires a systematic study of huge data from various parameters like land information, soil quality, soil pH, sources of irrigation, environmental data, etc. (Dahikar et.al. 2014). As prediction of the crop yield deals with a large set of databases, making this prediction system a perfect dataset for applying data mining methodologies that majorly helps acquire knowledge to achieve higher crop yields (Sudha et.al. 2018). With the advent of data mining techniques and big data, crop yield can be predicted by deriving useful information from the agricultural data that helps farmers to decide on the crop they would like to plant in future leading to maximum productivity and profit (Majumdar et.al. 2017; Grassinia et.al. 2015).

Everingham et.al.(2015b) reported the benefits of a modern data mining method over contemporary, time-honored methods like stepwise linear regression modeling. These authors used a random forest modeling method (Breiman, 2001) to investigate how climate attributes relate to sugarcane productivity in Australia. The random forest technique's key advantage is that it can investigate nonlinear relationships between the predictors and the response variable using an ensemble learning approach. Ensemble methods that involve making multiple attempts from different data or models to predict a response and as a result, the robustness and accuracy of predictions is increases when compared to using any one data set or model (Breiman, 2001; Everingham et.al. 2009), linear regression (Craig &Huettmann 2009), and linear discriminant analysis approaches (Everingham et.al. 2007b; Gromski et.al. 2014) and consequently, random forests have been applied in several agricultural-related issues.

Random forests have been used for the prediction of yields for various crops across Canada and USA and to identify important variables having a high impact for yields (Fukuda et.al. 2013; Newlands et.al. 2014; Tulbure et.al. 2012). Few researchers used Random forest analysis of Big Data sets to investigate other vital problems in agriculture (Philibert et.al. 2013; Abdel-Rahman et.al. 2013). Even the random forest regression model is also highly helpful in retrieving the valuable data from the available, poor quality, less rigorous data sources, or if the data is not available (Sagar & Cauvery, 2018).

From the above review, primarily three research themes have emerged. Firstly, the reviewed literature highlights that in developing countries, including India, most of the issues and problems faced by the farmers are region-specific. These are primarily due to changing climatic conditions, soil characteristics and constraints, land problems, unavailability of quality water, lack of marketing advancement, infrastructure deficiency, shortage of capital, weakness of supporting services, lack of training and education of farmers. Secondly, most of the studies indicated that the data mining approaches such as random forest algorithm is useful in predicting crop yield for a particular region. Even if any method has already been adopted and implemented to handle Big Data in various distributed environments, many questions remain open. Finally, in NER, India farmers are more stressed in producing higher crop yields due to the influence of unpredictable environmental changes and significant reduction of water resources (Foodstart, 2014), and therefore predicting the crop yield well in advance before its harvest can help the farmers and government organizations to make appropriate planning like selling, storing, fixing minimum support price, importing/exporting etc (Joshi et.al. 2018). The yield prediction model may provide a unique opportunity to overcome challenges and improve crop yield gap prediction ideology.

Research Methods

Initially, a questionnaire to be designed based on an initial set of chosen variables from the secondary literature and then it would be distributed to the representative sample. The raw data would be stored in an excel file for further analysis after cleaning and pre-processing and eventually would be divided into training and test data. Then with training data, a proposed prediction model would be configured with a default setting of random forest package of R statistical software, and subsequently the model would be tuned with hyper-parameters for improved accuracy. Finally, with the testing data, the model would be evaluated, and the accuracy of the model would justify the validity of the proposed prediction model. This model would also highlight the top twenty (20) variables that would significantly impact crop yields in hilly areas of NER, India.

Data Collection:

This study was conducted in the hilly areas of the North-eastern regions of India. This region is selected considering that it occupies a significant place in India's plan for economic development both for socio-economic and geo-political reasons It is also observed that the NER, India has often been visualized as the remote, landlocked backward region of a dynamic economy and the relatively low development indicators. The lack of adequate information about the new or advanced technology along with high poverty rates, impacts on the overall progress of the region (Chulet et.al. 2017).

A representative sample was arrived through multistage random sampling out of the eight (8) North Eastern States. Initially, one state is selected randomly, and then in the next stage, three districts comprising of 16 blocks are chosen from the selected state. Finally, to arrive at the sample, an equal number of 7(seven) agricultural experts as respondents are considered from each block through convenience sampling.

A total of 112 no. of questionnaires were distributed to the sample respondents and out of which 96 nos. of filled questionnaires were received with a valid response rate. The questionnaire was designed based on an initial set of chosen variables based on their relevance and contribution to the agricultural field. The first section of the questionnaire represents the respondents' demographic data, which are used as a control variable. The data are pertaining to the respondent's (agricultural expert) gender, age, education, working experience, job level, occupation details are collected. The second section of the questionnaire represents the key factors responsible for high crop yield. Besides, few variables are also added considering the limitation of the North Eastern Region. The questionnaires are presented in five points Likert scale (5 for highly satisfied, 4 for satisfied, 3 for neutral i.e. neither satisfied nor dissatisfied, 2 for dissatisfied, and 1 for highly dissatisfied) and the reason for chosing Likert scale is that it tends to reach the upper limit of reliability (Nunnally, 1978). The items of the constructs in the questionnaire are considered after the acceptance of reliability and validity test with cronch alpha and exploratory factor analysis, respectively. Finally, these items are considered to build the proposed system. Then, collected data entered into an excel datasheet for further analysis.

Data Pre-processing

Initially, raw data is cleaned and missing values were removed and then the final dataset obtained is stored in CSV format of Ms Excel application for reading into R software. The dataset contained of seventy-three predictor variables and one response variable (yield) as shown in the annexure-A, and most of these attributes are categorical. The response variable consisted of five levels such as High (H), Low (L), Neutral (N), Very High (VH), and Very Low (VL).

Reading and partitioning the data

The collected dataset read and viewed in the R environment and then partitioned into two datasets. One partition specifies the data that is to be evaluated, while the other refers to the set of data concerning which the test is carried out. In this analysis, the partition is done by 70/30 and 70% of the records are used for training, i.e. 336 number of records and the rest 30% is used for testing, i.e. 144 number of records. The analysis is carried out on the test data set, and based on the result, the accuracy of the proposed crop prediction model is evaluated.

Building Proposed prediction model with Random Forest Algorithm

The free statistical software R (Besse& Villa-Vialaneix, 2014) is de facto the es-peranto in the statistical community and consists of flexible and widely used programs for designing random forests. The caret and random forest (R.F) package of R software are the R.F. algorithm using Breiman and Cutler's Fortran code, contains many options together with detailed documentation. For building proposed system, we have selected it for classification of the levels of crop yield. To build the model, we have first loaded the pre-processed data to the working directory of the R workspace and then the RF packages are used to predict the crop yield.

Mechanics of the Random Forest Algorithm

The main feature of this method is to use different datasets for building a tree, which is achieved by a method called bootstrap aggregating (bagging).

Our dataset is of size 480. From this dataset, since the bootstrap aggregated sample is created by sampling with replacement, some data points will not be selected anytime. Generally, on average, each sample will use about

two-thirds of the available data points, and one-third data points will not be chosen in any samples, so the model will not be trained on those 1/3rd data points.

Using subsets of predictor variables

Bootstrap aggregating (bagging) reduces over-fitting to a certain extent, but it does not eliminate overfitting issues. This is because specific input predictors influence the tree split, and they overshadow weak predictors. These predictors play an essential role in the early split of the decision tree, and subsequently, they influence the size and structures in the forest. The random forest selects a random set of subset predictors for each split to be identical. So strong predictors cannot overshadow other fields, and hence we get more diverse forests.

In random forest classification trees, the splitting decision is based on the following methods:

Gini Index - It is a measure of node purity. If the Gini index takes on a smaller value, itindicates that the node is pure. For a split to occur, the Gini index for a child node must be less than that parent node.

Entropy - Entropy is a measure of node impurity. For binary class (a, b), entropy is maximum at p = 0.5. The entropy is minimum when probability either 0 or 1.

Entropy = - p(a)* log(p(a)) - p(b)*log(p(b)).

In a nutshell, every tree attempts to generate rules so that the resultant terminal nodes could be as pure as possible-the higher the purity, the lesser the uncertainty to make the decision.

The model with default setting:

Initially, dataset is trained with caret package in R with default setting to evaluate the crop yield as shown in table 1:

Table 1: Model code default settings

```
set.seed(1234)
# Run the modelrf_default<- train(Yield~ data=trainDF2, method = "rf", metric = "Accuracy", trControl = trControl)
```

Table 2 shows the output of the random forest model, and the accuracy of the model is 86%. However, the accuracy of the model can be further improved by the tuning process.

Tuning Random forest Model

This step is carried out to improve the model's accuracy and selecting the value that provides the least error. However, the tuning the hyper parameters for a model is cumbersome work. There can be many complex permutations and combinations for a set of hyper-parameters. Trying all combinations can be a time and memory-consuming and a better approach could be that the algorithm decides the best set of hyper parameters. There are two standard methods for tuning: a) grid search and b) random search. A Random search does not evaluate all the combinations of hyper-parameters. Instead, it will randomly choose any combination at every iteration. The advantage is it lowers the computational cost, memory cost, and less time required. Therefore, we have selected random search method for tuning the model

Table 2: Output of Random Forest Model with default setting

```
Output
Random Forest
335 samples
68 predictor
5 classes: 'H', 'L', 'N', 'VH', 'VL'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 301, 301, 301, 304, 302, 300…
Resampling results across tuning parameters:
Mtry  Accuracy   Kappa
  2   0.8683850  0.8353689
137   0.8503642  0.8128673
272   0.8475910  0.8094155
```

The hyper-parameters for tuning the model are ntree (number of trees), mtry (number of variables to sample), sample size (the number of samples to train on), node-size (minimum number of samples within the terminal nodes) and max-nodes (maximum number of terminal nodes).

The default algorithm, as shown in table 1, uses 500 no. of trees and tested with three different values of mtry such as: 2, 6 and 10. The final value used in the model was mtry = 2 with a maximum accuracy of 0.86. However, the higher score is achieved after tuning the model with mtry, ntree, node-size, and max-nodes through random search and values of these are; mtry = 2, node-size = 14, ntree = 2000, max-nodes =14.

Evaluating the Proposed model:

Finally, the test data set is used to evaluate the build model with predict function of R as follows:

Table 3: Code of Confusion Matrix

```
prediction <-predict(fit_rf, testDF2) confusion Matrix(prediction, testDF2$Yield)
```

The confusion matrix along with the accuracy score as the output of the predict function is shown in table 4.

Table 4: Result of the Confusion Matrix Model

```
Confusion Matrix and Statistics
          Reference
Prediction  H  L  N VH VL
         H 25  2  0  0  2
         L  1  2  1  0  1
         N  1  0  2  0  1
         VH 0  1  1 28  2
         VL 2  1  0  1 23
Overall Statistics
Accuracy : 0.8828
                 95% CI : (0.8189, 0.9302)
No Information Rate : 0.2
P-Value [Acc> NIR] :< 2.2e-16
Kappa : 0.8534
Mcnemar's Test P-Value : NA
Statistics by Class:
                     Class: H  Class: L Class: N Class: VH
Class: VL
Sensitivity          0.8621   0.8621  0.9310     0.9655
0.7931
Specificity          0.9655   0.9741  0.9828     0.9655
0.9655
PosPred Value        0.8621   0.8929  0.9310     0.8750
0.8519
NegPred Value        0.9655   0.9658  0.9828     0.9912
```

Table 4 shows that after tuning, the model achieved an overall accuracy of 0.88%, which is better accurate than default random forest model.

0.9492					
Prevalence	0.2000	0.2000	0.2000	0.2000	0.2000
Detection Rate	0.1724	0.1724	0.1862	0.1931	0.1586
Detection Prevalence	0.2000	0.1931	0.2000	0.2207	0.1862
Balanced Accuracy	0.9138	0.9181	0.9569	0.9655	0.8793

Identifying variable importance:

The importance of the variables affecting the crop yield can be determined with the mean decrease accuracy and the mean decrease in the Gini coefficient. It provides the idea about the factors that directly impact the crop yield measurement and how each variable contributes to the homogeneity of nodes and leaves. The result indicated that after tuning the Gini index score is higher, the variable's importance in the model and the node is pure, which is shown in figures 1 and 2.

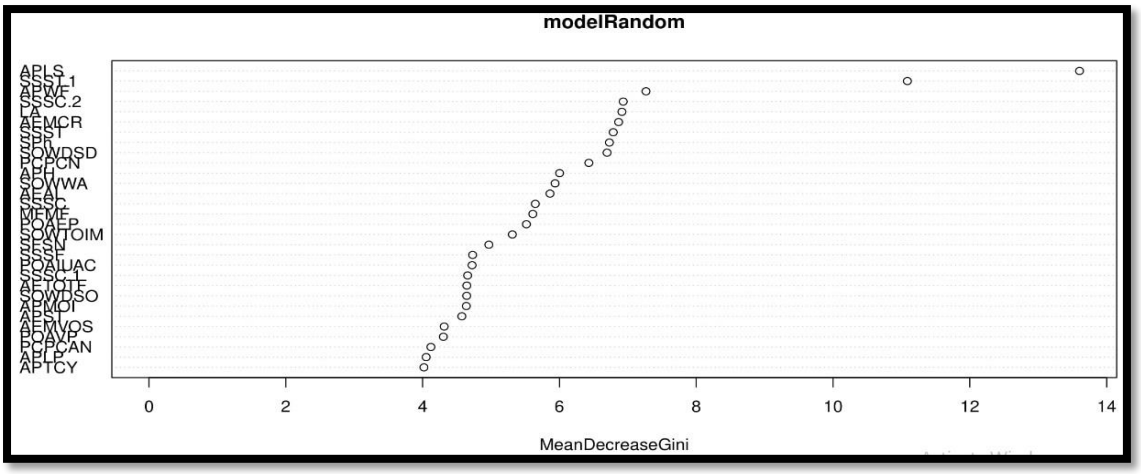

Fig 1: Decrease Gini before tune

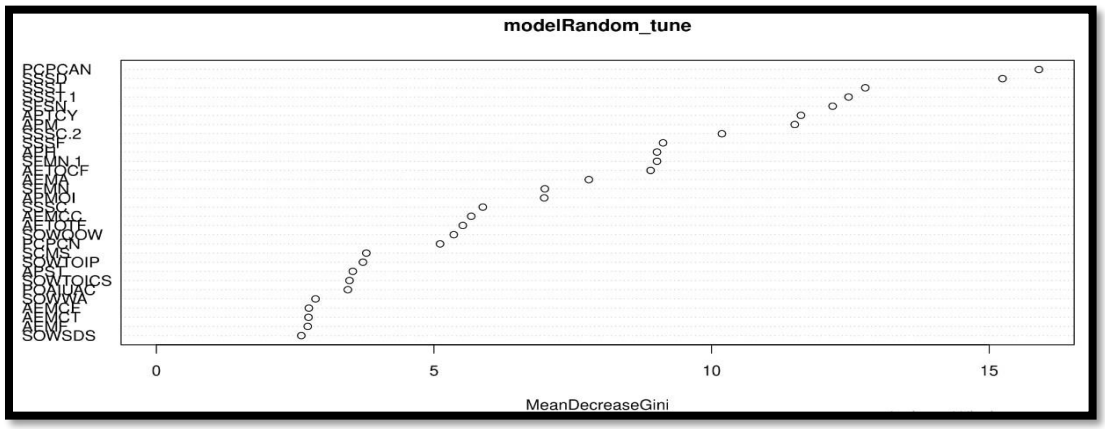

Fig 2: Decrease Gini after tune

The top 20 high importance variables obtained after tuning of the Random forest model is shown in figure 3 and it is observed that only natural measures of pest control have the highest effect on crop productivity and on the other hand, quality of water has the less impact on crop productivity as per the mean decrease in the Gini coefficient.

Besides, figure 3 depicts that in case of a mean decrease in accuracy, applied chemical measure of pest control also has the least effect on high crop yield, and this confirms that excessive use of chemicals may reduce crop productivity. Other important variables are agriculture production (traditional cropping i.e. APTCY, harvesting method i.e. APH, method of irrigation i.e. APMOI, manure frequency i.e. APM), pest control (applied natural i.e. PCPCAN and applied conventional i.e. PCPCN), soil suitability (soil density i.e. SSSD, soil composition i.e. SSSC.2, soil temperature i.e. SSST.1, soil texture i.e. SSST, soil fertility i.e. SSSF, soil color i.e. SSSC), soil fertility nutrient (secondary nutrient i.e. SFSN, minor nutrient i.e. SFSN.1), agriculture economics (types of cooperative farming i.e. AETOCF, market area i.e. AEMA, market competition i.e. AEMCC, types of traditional farming i.e. AETOTF), sources of water (quality of water i.e. SOWQOW).

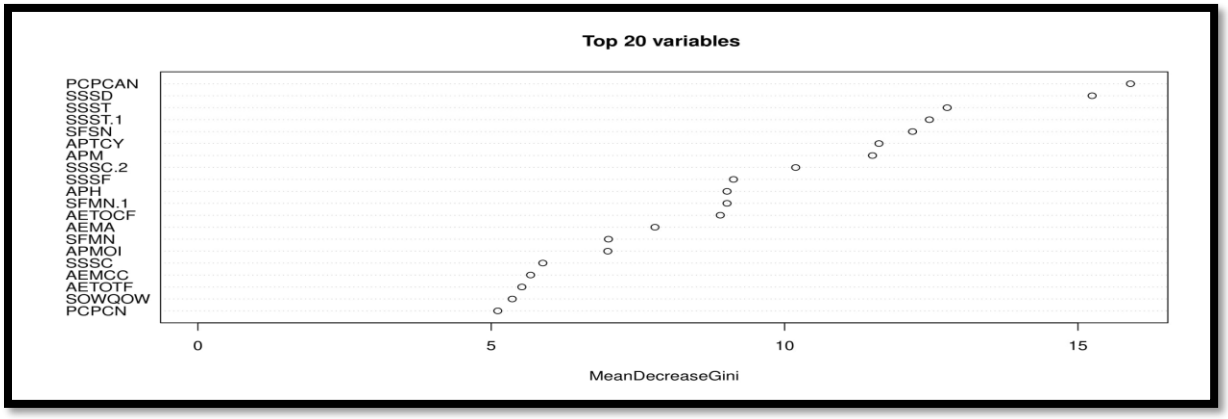

Fig 3: Top 20 important variables after tune

The models were rebuilt with the most critical predictor variables and good accuracy of 0.88% and minor OOB error, as shown in table 4. This model offers the opportunity to improve performance while keeping the number of predictor variables low to minimize the risk of over-fitting.

Results and Discussion

The variable importance measures of the random forest algorithm predicted the top twenty (20) important variables that are impacting high crop yield and revealed that pest control (natural) was the most influential variable, followed by soil suitability (soil density, soil composition, soil temperature, soil texture, soil fertility, soil color), soil fertility nutrient (secondary nutrient, minor nutrient), agriculture production, agriculture economics (types of co-operative farming, market area, market competition, types of traditional farming) and sources of water (quality of water).

The southwest monsoon is responsible for 90 percent of the annual rainfall to the region and being hilly or mountainous areas, and the temperature is generally lower than the plains and these help crops to grow faster, increasing its yield (Barah, 2005). However, these are also congenial for increasing crop pests and diseases (Deka et.al. 2010) and eventually may consistently reduce crop yield. Therefore, natural pest control measures may reduce pests and disease, thereby increasing crop yield manifold, which justifies the findings of the present study. Further, the North-eastern region is also known for using organic manures such as cow dung, neem seed powder coating, bio-fertilizers, organic matter recycling, enrichment of compost, vermicomposting, animal manures, urine, farmyard manure, litter composting, use of botanicals, and green manuring. Thus, the application of these integrated natural pests management in farmers' fields will consistently increase crop yield. The finding of the

study is in line with the finding of Rahman et.al. (2009). According to earlier studies, in the hilly areas of NER, India, mostly soil are sandy clay loam and red lateritic soils with high water retention and more pH level. In addition, the soil temperature is also low as compared to other plain areas (Barah, 2005). These results in increasing the fertility of the soil (Deka & Sarmah, 2010). Thus, these finding is similar to the findings of the present study, which also confirms that soil types, texture or temperature have a positive effect on crop yield. Further, the present study also shows that organic matter content and N, P, K content in the soil has a strong impact on crop production and high yield, and this finding is consistent with the finding of Roy et.al. (2014a), which also supported that the high proportion of nitrogen and organic matters and medium to low available potassium levels in the soils of NER, India supports crop production. According to Dhas et.al.(2006), NER, India mostly consists of jhum land, and the use of traditional means of planting or cropping patterns such as crop diversification, co-operative farming, and mixed farming helps in the production of forest and agricultural crops. This finding is also in sequence with the present study, which shows that co-operative farming, traditional farming, and traditional cropping system supports in optimizing and improving the field for higher yield. Further, the present study confirms that market area and competition favor the high crop yield. However, North East India, being the backward region, the market competition is very less, and the region provides opportunities to widening the market for agricultural products. The earlier study of Banerjee (2006) also highlighted that there are competition and greater demand for few products like spices, medicinal and aromatic plants in the region. In addition, the present study indicated that methods of irrigation and quality of water also influence crop production. But it is observed that the farmers in the region are dependent on other sources of water that contain salts and minerals despite the highest rainfall in the region due to the infrastructural deficiency. Subsequently, the present study developed a crop prediction model using a random forest algorithm with 88% accuracy on the test data set, and this confirms that the random forest model can accurately estimate crop production with some error. This finding is also in consistent with previous studies showing that the random forest model can accurately estimate the yield of sugarcane and mango fruit yields (Fukuda et.al. 2013; Everingham et.al. 2015b).

Conclusion:

In India, crop yield is season-dependent and majorly influenced by a particular crop's economic and biological factors. To achieve sustainable amplification of agricultural yields, it is necessary that the farmers are aware and informed about the yield gaps. The paper attempts to examine if a data mining approach with growing data could offer new insights that can explain crop productivity in the hilly region of North East India. The identified key factors responsible for high crop yield would help decision-makers and policymakers conceptualize and visualize farmers' growth. These factors may boost farmers to realize the state's farming potential and subsequently influence their decision to continue farm activity. The same can be applied to a single crop or to various other crops, and eventually, corrective measures can be undertaken to increase the crop yield in different regions of India.

In the absence of real online data such as big data, it must be stressed that the data mining approach that put forth the proposed system also has some limitations which can be considered as a future enhancement, and the identified key factors can be combined with new theories and models for further research. Further, a database is a serious constraint to effective policy analysis in the agricultural economy in the region. Therefore, an agricultural database must be streamlined properly on a priority basis by taking the help of the electronic revolution.

Reference:

Abdel-Rahman, E.M., Ahmed, F.B., & Ismail, R. (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing*, 34(2), 712-728. Doi:10.1080/01431161.2012.713142.

Aggarwal, P.K. (2009). *Determinants of Crop Growth and Yield in a Changing Climate*, ed. Wani, S.P. CAB International: Rainfed Agriculture: Unlocking the Potential, 73-80.

Aggarwal, P.K., Talukdar, K.K. & Mall, R.K. (2000). Potential yields of rice–wheat system in the Indo-Gangetic plains of India. New Delhi, India: Rice–Wheat Consortium Paper Series, RWCIGP, CIMMYT, 10.

Aker, J. C. (2011). Dial A for agriculture: A review of information and communication technologies for agricultural extension in developing countries. *Agricultural Economics Journal*, 42(6), 631–647. Doi.org/10.1111/j.1574-0862.2011.00545.x.

Ali, A.N.M & Das, I. (2017). Tribal Situation in North East India. *Studies of tribe and tribals journals*, 1(2), 1410148. Doi:10.1080/0972639X.2003.11886492.

Anantharaman, M., Mantri, V., Shanpru, E., Sohliya, B.K., Rathore, N.S., Suting, A., Prain, G., & Bertuso, A. (2016). *Food Resilience through Root and Tuber Crops in Upland and Coastal Communities of the Asia-Pacific (FoodSTART).* Meghalaya, India: Roots and Tubers for Livelihood Enhancement in Meghalaya Study under IBDLP Meghalaya Publication, 1-87.

Arun, K. & Sharma. 2006. *A Hand Book of Organic Farming*. Jodhpur: India, Agrobios.

Banerjee, A., (2006). *Economic Growth and Sustainability of North Eastern States*, Passah, P.M. (Ed.). New Delhi, India: Defence of Regional Economic Development in: A Case for the North East, Akansha Publishing House.

Barah, B.C. (2005). *Prioritisation of Strategies for Agricultural Development in the North-eastern India*. New Delhi: Proceeding Series No. NCAP.

Barah, B.C. (2007). Strategies for Agricultural Development in the North-East India- Challenges and Emerging Opportunities. *66th Annual Conference of the* Indian *Society of* Agricultural Economics *Journal*, 62(1), 1-19.

Bernardo, J. M., & Smith, A. F. M. (2001). *Bayesian Theory. Measurement Science and Technology*, 12, 211.

Besse, P. & Villa-Vialaneix, N. (2014). *Statistic big data analytics*. Retrieve from http://arxiv.org/abs/1405.6676.

Bhanja, S. N., Mukherjee, A., Rodell, M., Wada, Y., Chattopadhyay, S., Velicogna, I., &Famiglietti, J. S. (2017). Ground water rejuvenation in parts of India influenced by water policy change implementation. *Scientific Reports Journal*, 7(1), 7453. DOI: 10.1038/s41598-017-07058-2.

Bharadi, V. A., Joshi, A. M. &Patade, S. S.(2017). Analysis and Prediction in Agriculture Data using Data Mining techniques. *International Journal of Research In Science & Engineering.*

Birthal, P.S., Jha, A.K., Joshi, P.K., & Singh, D.K. (2006). Agricultural Diversification in North Eastern Region of India: Implications for Growth and Equity. *Indian Journal of Agricultural Economics*, 61(3), 1-13. DOI: 10.22004/ag.econ.204466.

Borthakur, D.N. (2002). *Shifting Cultivation in Northeast India: An Approach towards Control*, Deb, B.J. (ed.). New Delhi: Development Priorities in Northeast India, Concept Publishing Company.

Breiman, L. (2001). *Random forests, Machine Learning*. Springer Publication, 45 (1), 5–32. DOI: 10.1023/ A: 1010933404324.

Retrieved from http://www.springerlink.com/content/u0p06167n6173512/ fulltext.pdf.

Chandrasekaran, B., Annadurai, K., Somasundaram, E.(2010). *A textbook on agronomy*. New Delhi, India: New age International (p) limited publishers, 1-856.

Chulet, H., Anantharaman, M., Shanpru, E., &Prain, G.(2017). *Potato Production, Marketing, and Utilization in Meghalaya, India- Results of a Value Chain Assessment. Food Resilience through Root and Tuber Crops in Upland and Coastal Communities of the Asia-Pacific (FoodSTART+).* Laguna, Philippines: International Potato Center (CIP), Research Brief No. 8, 4. Retrieve from https://hdl.handle.net/10568/92897.

Craig, E., &Huettmann, F. (2009). *Using blackbox algorithms such as TreeNET and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: An overview and example using G*,Wang H (ed) Intelligent data analysis: developing new methodologies through pattern discovery and recovery. Information Science Reference, Hershey, 65–84. DOI:10.4018/978-1-59904-982-3.ch004.

Dabral, P.P. (2002). Indigenous techniques of soil and water conservation in North-Eastern Region of India. *Beijing: 12th ISCO Conference*, 90- 96.

Dahikar, S., & Rode, S.V. (2014). Agricultural crop yield prediction using Artificial-Neural Network Approach. *International Journal of Innovative Research in Electrical, Electronic, Instrumentation and Control Engineering (IJIREEICE*), 2(1), 683-686.

Das, A., Layek, J., Ramkrushna, G,.&Babu, S. (2018). Integrated Organic Farming System in North East India. *Conservation Agriculture for Advancing Food Security in Changing Climate*, 1, 301-318.

Das, T.K., Samajdar, T., Mokidul, I., Singh, N. A., &Marak, G. (2016). Traditional farming system: a case study of garo tribe in west garo hills district of Meghalaya, North-Eastern India. *International Journal of Agriculture Sciences*, 8(*50*), 2140-2145, Bioinfopublication.

Deb, D.L. (1994). *Natural Resources Management for Sustainable Agriculture and Environment.* New Delhi: Angkor publishers Limited.

Deka, B.C., Nath, A., Jha, A.K., Patel, R.K., Yadav, R.K., Singh, A., Kumar, R., &Ngachan, S.V. (2010). *Package of Practices for Horticultural crops of NEH Region*. ICAR Research Complex for NEH Region, Umiam (Meghalaya).

Deka, P.K. &Sarmah, D. (2010). Shifting cultivation and its effects in regarding of perspective in Northern India. *International Journal of Commerce and Business Management*, 3(2), 157-165.

Deshmukh, N.A., Patel, R.K., Verma, V.K., Firke, D.M., &Jha, A.K. (2017). Potential Fruits and Plantation Crops of Meghalaya. *Horticulture for Economic Prosperity and Nutritional Security in 21st Century*, 225-241.

Dev, B.J. & Datta, B..R. (2006). *Changing Agricultural Scenario in North East India*. New Delhi: Concept Publishing Company.

Dhas, S.R., Jeeva, N., Laloo, R.C., & Mishra, B.P.(2006). Traditional agricultural practices in meghalaya, North east India. *Indian Journal of Traditional Knowledge*, 5(1), 7-18.

Dhivya, B. H., Manjula, R., Siva, B.S, &Madhumathi, R. (2017). A Survey on Crop Yield Prediction based on Agricultural Data, International Journal of Innovative Research in Science, Engineering and Technology, 6(3).

Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, Springer Verlag/EDP Sciences/INRA, 36 (2), 27. DOI: 10.1007/s13593-016-0364-zff.

Everingham, Y.L., Inman-Bamber, N.G., Sexton, J., Stokes, C. (2015a). A dual ensemble agroclimate modelling procedure to assess climate change impacts on sugarcane production in Australia. *Agricultural Science Journal*. DOI:10.4236/as.2015.68084.

Everingham, Y.L., Lowe, K.H., Donald, D.A., Coomans, D.H., Markley, J. (2007b). Advanced satellite imagery to classify sugarcane crop characteristics. *Agronomics Sustainability Development*. DOI: 10.1051/agro:2006034.

Everingham, Y.L., Sexton, J., Robson, A. (2015b) *A statistical approach for identifying important climatic influences on sugarcane yields*. In: ProcAustSoc Sugar Cane Technol. Bundaberg, Australia, 8–15.

Everingham, Y.L., Smyth, C.W., Inman-Bamber, N.G.(2009). Ensemble data mining approaches to forecast regional sugarcane crop production. *Agricultural for Meteorology*. DOI:10.1016/j.agrformet.2008.10.018.

Foodstart (2014). *Report of the assessment phase for the FoodSTART project Root and Tubers for Food Security in Asia Pacific Focus Sites in India*. CIP, New Delhi.

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., & Müller, J. (2013). Random Forests modelling for the estimation of mango (Mangiferaindica L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural Water Management Journal*, 116 (39), 142-150. Doi:10.1016/j.agwat.2012.07.003.

Grassinia, P., Lenny, G.J. Bussel, V., Warta, J.V., Wolf, J., Claessens, L., Yanga, H., Boogaarde, H., Groote, H., Martin, K., Ittersumb, V., Kenneth, G.C. (2015). How good is good enough: Data requirements for reliable crop yield simulations and yield-gap analysis. *Field Crops Research*, 49–63.

Grisso, R. D., Alley, M. M., McClellan, P., Brann, D. E., & Donohue, S. J. (2009). *Precision farming: a comprehensive approach*.

Gromski, P.S., Xu, Y., Correa, E., Ellis, D.I., Turner, M.L., Goodacre, R. (2014). A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Anal Chim Acta*. DOI:10.1016/j.aca.2014.03.039.

Hazarika, U.K., Munda, G.C., Bujarbaruah, K.M., Das, A., Patel, D.P., Prasad, Kamta, Rajesh Kumar.,Panwar, A.S., Tomar, J.M.S., Bordoloi, J., Sharma, M., and Gogoi, G. (2006). *Nutrient management in organic farming*. Technical Bulletin no.30. ICAR Research Complex for NEH Region, Umiam, Meghalaya.

ICAR, (2013). Package of practices for organic production of important crops in NEH region. Published by ICAR Research Complex for NEH Region, Umiam – 793 103, Meghalaya.

ISPS, (2005). *Experience in Collaboration-Ginger Pests and Diseases*. Inter-cooperation India Programme, Series 1, (Indo-Swiss Project-Sikkim, Hyderabad, India), 57.

Jawad, F., Choudhury, T.R., Sazed, S.M., Yasmin, S., Rishva, K.I., Tamanna, F., Rahman, R.M. (2016). *Analysis of Optimum Crop Cultivation Using FuzzySystem*. Dhaka, Bangladesh: North South University,

Joshi, H., Gawade, M., Ganu, M., Porwal, P. (2018). Crop Yield Prediction Using Supervised Machine Learning Algorithm. *IOSR Journal of Engineering,* 2278-8719, 35-42.

Kannan, Elumalai, & Sundaram, S. (2011). *Analysis of trends in India's agricultural growth, The Institute for Social and Economic Change*. Bangalore, India, Working paper 276.

Kumar, P. (1998). *Food demand and supply projection for India*. New Delhi, India: Agricultural Economics Policy Indian Agricultural Research Institute, 98–01,

Kumar, R., Singh, M.P., Kumar, P., and Singh, J.P. (2015). Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique. *International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*.

Kumaraswamy, S., & Shetty, P.K. (2016). Critical abiotic factors affecting implementation of technological innovations in rice and wheat production: A review. *Agricultural Reviews*, 37(4), 268-278. DOI: 10.18805/ag.v37i4.645.

Lee, D.R., (2005). Agricultural sustainability and technology adoption: Issues and policies for developing countries. *American Journal of Agricultural Economics*, 87, 1325-1333. DOI: 10.1111/j.1467-8276.2005.00826.x.

Li, X. (2008). Research on Classification Calculation Way of a Great Amount of Data. *Database Sampling Computer Science*, 35(6), 299.

Li, X., Chuan, T., Xin, F.(2010). *Research on Classification Technology in Data Mining Modern Electronics Technique*, 33(20), 86-8.

Liliane, T.N. & Charles, M.S. (2020). Factors Affecting Yield of Crops. *Intechopen Publication*. DOI: http://dx.doi.org/10.5772/intechopen.90672.

Liu H., Chen, J., Chen, G. (2002). Review of Classification Algorithms in Data Mining. *Journal of Tsinghua University (Science and Technology)*, 42(6), 727-30.

Liu, B., Ma, Y., & Wong, C.K. (2001). *Classification using association rules: weakness and enhancements*. In Vipin Kumar, et al. Data Mining for Scientific Applications.

Majumdar, J., Naraseeyappa, S., &Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques-application of big data. *Big data open access Journal*, 4, 20(2017). DOI:10.1186/s40537-017-0077-4.

Manjula, S., Devi, Y. V., &Thangamani, M. (2016). Predictive Analysis of Rainfall Data to Help the Farmers. *International Journal of Advanced Research in Computer Science & software Engineering*, 6(3), 1-20.

Mythili, G., & Goedecke, J. (2016*). Economics of land degradation in India*. In E. Nkonya, A. Mirzabaev, & J. von Braun (Eds.), Economics of land degradation and mprovement—A global assessment for sustainable development. Cham: Springer Journal Publication, 431–469. Doi.org/10.1007/978-3-319-19168-3_15.

Narkhede, U. P., & Adhiya, K. P.(2014). Evaluation of Modified K-Means Clustering Algorithm in Crop Prediction. *International Journal of Advanced Computer Research.*

Neenu, S., Biswas, A.K., & Rao, A.S. (2013).Impact of climatic factors on crop production - a review. *Agricultural Reviews Journal*, 34 (2), 97-106.

Newlands, N.K., Zamar, D.S., Kouadio, L.A., Zhang, Y., Chipanshi, A., Potgieter, A., Toure, S., Hill, H.S.J. (2014). An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Front Environmental Science Journal*. DOI:10.3389/fenvs.2014. 00017.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Patangray, A.J., Patil, N.G., Singh, S.K., Tiwari, P., Mishra, V. N., Pagdhune, A.R. & Patil, B.A. (2016). Soil Suitability Evaluation of Major Crops for Sustainable Land Use Planning in Kupti Watershed, Yavatmal District, Maharashtra. *Agropedology Journal*, 26 (02), 117-131.

Patel, R.K., Singh, A., Yadav, D.S., & De, L.C. (2008). *Underutilized fruits of North-Eastern region, India*. Underutilized and under exploited Horticultural crops, K.V. Peter (Ed.) New India Publishing Agency, 4, 223-238.

Philibert, A., Loyce, C., &Makowski, D. (2013). Prediction of N2O emission from local information with Random Forest. *Environmental Pollution Journal*. DOI:10. 1016/j.envpol.2013.02.019.

Pingali, P. (2010). *Agriculture renaissance: Making agriculture for development work in the 21st century*. In P. Pingali& R. Evenson (Eds.), Handbook of agricultural economics (3867–3894). Elsevier publication. Retrieved from http://www.sciencedirect.com/science/article/pii/S1574007209040742.

Pingali, P., Aiyar, A., Abraham, M., & Rahman, A. (2019). Agricultural Technology for Increasing Competitiveness of Small Holders. *Transforming Food Systems for a Rising India*, 215-240.

Prodhan, A.Z.M.S., Islam, M.S., & Islam, M.M.(2018). Effect of soil and environment on winter vegetables production. *MOJ Food Processing Technology*, 6(4), 384–389. DOI: 10.15406/mojfpt.2018.06.00192.

Rahman, H., Bujarbaruah, K.M., Srivastava, L.S., Karuppaiyan, R., Avasthe, R.K., & Singh, M. (2007). *Status of Ginger Cultivation in Sikkim with Special reference to Disease Management*, In: DBT Interactive Workshop for the Generation of Network Programme on Management of Ginger Diseases and Pests in Northeastern and Himalayan Region. ICAR Research Complex publication for NEH Region, Sikkim Centre, Gangtok, 5-6, 35-47.

Rahman, H., Karuppaiyan, R., Kishore, K., &Denzongpa, R. (2009). Traditional practices of ginger cultivation in Northeast India. *Indian Journal of Traditional Knowledge*, 8(1), 23-28.

Raman, S., Devineni, N., & Fishman, R. (2015). Can improved agricultural water use efficiency save India's groundwater. *Environmental Research Letters*, 10(8), 84022. Doi.org/10.1088/1748-9326/10/8/084022

Raza, A., Razzaq, A., Mehmood, S.S., Zou, X., Zhang, X., &Lv, Y. (2019). Impact of climate change on crop adaptation and strategies to tackle its outcome: A review. *Journal of Plants Journal*, 8(2), 1-34. DOI: 10.3390/plants8020034.

Roy, A., Dkhar, D.S., Tripathi, A.K., Singh, N. U., Kumar, D., Dasand, S.K., &Debnath, A. (2014b). Growth Performance of Agriculture and Allied Sectors in the North East India. *Journal of Economic Affairs*, 59, 783-795.

Roy, A., Tripathy, A.K., Mohanty, A.K., Singh, N.U., Dhar, D.S., Baruah, S., Saxena, R., Ngachan, S.V. (2014a). *Commodity profile on potato, tomato, ginger, turmeric, and pineapple in Meghalaya*. Umiam, Meghalaya: ICAR Research Complex for NEH Region.

Sagar, B.M., & Cauvery, N. K. (2018). Agriculture Data Analytics in Crop Yield Estimation: A Critical Review. *Indonesian Journal of Electrical Engineering and Computer Science*, 12(3), 1087-1093. DOI: 10.11591/ijeecs.v12.i3.pp1087-1093.

Saussure, S., Plantegenest, M., Thibord, J.B., Larroudé, P., Poggi, S.(2015). Management of wireworm damage in maize fields using new, landscape-scale strategies. *Agronomics Sustainable Development*. DOI: 10.1007/ s13593-014-0279-5.

Shah, F., & Wu, W. (2019). Soil and crop management strategies to ensure higher crop productivity within sustainable environments. *Journal of Sustainability*, 11(1485), 1-19. DOI: 10.3390/ su11051485.

Shang, Q., Ling, N., Feng, X., Yang, X., Wu, P., Zou, J. (2014). Soil fertility and its significance to crop productivity and sustainability in typical agro ecosystem: A summary of long-term fertilizer experiments in China. *Plant and Soil Journal*, 381, 13-23. DOI: 10.1007/ s11104-014-2089-6.

Shettar, A. A., &Angadi, S. A. (2016). Efficient Data Mining Algorithm for Agriculture Data. *International Journal of Recent Trends in Engineering and Research*.

Singh, R.K. (2002). *Soil conservation measures in agricultural land, in Integrated Watershed Management for Sustainable Development*. ICAR Research Complex publication for NEH Region, Umiam, Meghalaya, 104.

Singh, S.S. (2015). *Handbook of agricultural sciences*. India: Kalyani Publisher, 1-30.

Skocaj, D.M., & Everingham, Y.L. (2014). Identifying climate variables having the greatest influence on sugarcane yields in the Tully mill area. *In: Proceeding Australia Soc Sugar Cane Technology*, 53–61.

Sudha, V., Mohan, S., & Arivalagan, S. (2018). Big Data Analytics to Increase the Agricultural Yield by Using Machine Learning Approaches. *Asian Journal of Computer Science and Technology*, 7(1), 82-86.

Vetter, S. H., Sapkota, T. B., Hillier, J., Stirling, C. M., Macdiarmid, J. I., Aleksandrowicz, L., & Smith, P. (2017). Greenhouse gas emissions from agricultural food production to supply Indian diets: Implications for climate change mitigation. *Agriculture, Ecosystems & Environment*, 237(Suppl. C), 234–241. DOI: 10.1016/j.agee.2016.12.024.

Wang, C.H. (2014). Farming methods effects on the soil fertility and crop production under a rice vegetables cropping sequences. *Journal of Plant Nutrition*, 37, 1498-1513. DOI: 10.1080/01904167.2014.881876.

Zollingrand, B., &Krannich, R. (2002). Factors Influencing Farmers' Expectations to Sell Agricultural Land for Non-Agricultural Uses. *Rural Sociology*, 67(3), 442-463.

Annexure-A:

List of variables considered for the present study

Sr. No.	Category of variables	Category	Categorical Variables
1.	Predictors	Crop-Topography	Land area usage, land situations, method of sowing, land preparation, harvesting, varieties, traditional cropping system, seed treatment, method of irrigation, harvesting, varieties, weeding frequency, manure frequency, fertilizer frequency.
2.		Crop-Climate	Weather and time of sowing.
3.		Crop-Soil suitability	Quality of soil, soil texture, soil structure, soil density, soil temperature, soil fertility, soil colour, soil composition, soil depth, soil stickiness, soil plasticity, moist soil, dry soil, soil pH, major nutrient, minor nutrient, fertilizer usage, manure usage.
4.		Crop-Irrigation	Quality of water, type of irrigation size, irrigation canal, irrigation management, irrigation purpose, irrigation sources of water, irrigation conveyance system, surface irrigation system, surface drainage system, drainage ownership, drainage duration.

5.		Crop-Pest control	Pest control (PC) natural, PC applied conventional, PC applied modern, PC integrated pest management, protection on chemicals.
6		Crop-agricultural economics	Nature of agricultural production, agricultural finance, market seller, market time, market competition, market regulation, market area, market function, market volume of sale, fluctuation in agricultural price
7.		Crop-sustainability strategy	Types of traditional farming, system of farming, types of cooperative farming, agricultural labour, elements of sustainability, conservative farming, agricultural sustainability,
8.		Crop Problems	Land problems, irrigation problems, indiscriminate use of agro chemicals, vulnerability problems, Environmental problems and farmer's attitude.
9	Response	Yield	High (H), Low (L), Neutral (N), Very High (VH), and Very Low (VL).