ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach

Vaneesbeer Singh
Department of Computer Science and IT
University of Jammu
Jammu, India

Abid Sarwar
Department of Computer Science and IT
University of Jammu
Jammu, India

Vinod Sharma
Department of Computer Science and IT
University of Jammu
Jammu, India

Abstract: Agriculture sector is backbone of Indian Economy .However Agriculture sector in India is facing severe problem of maximising the crop productivity. Farmers lack in basic knowledge of nutrient content of soil, selection of crop best suited for soil and they also lack in efficient method of prediction of crop well in advance so that appropriate methods can be used to improve crop productivity and to make arrangements for storage, marketing well before harvest. This work presents an approach which uses different Machine Learning techniques in order to predict the category of the yield based on macro-nutrients and micro- nutrients status in dataset. The dataset considered for the crop yield prediction was obtained from Krishi Bhawan (Talab-Tillo) Jammu. The parameters present in the data are Macro-Nutrients (ph,Oc,Ec,N,P,K,S) and Micro Nutrients(Zn,Fe,Mn,Cu) present in samples collected from different regions of Jammu District .After analysis Machine learning algorithms are applied to predict the category of yield . The category, thus predicted will specify the yield of crops. The problem of predicting the crop yield is formulated as Classification where different classifier algorithms are used.

Keywords**:** soil, crop, yield, machine, learning, approach

1. INTRODUCTION

Agriculture sector is backbone of Indian Economy. More than half of population is dependent on agriculture. However farmers still lack in basic knowledge about their soils, which crop to sow in what type of soil, the efficient use of fertilizers and in addition to it they also lack in advanced techniques of prediction of crop yield at the time of sowing. Prior prediction of crop yield can be helpful in early determination of factors leading to decline in production.it can also be helpful in determine crop diseases ,proper use of pesticides, proper selection of different varieties of crop and in addition to it early prediction gives farmers an option for prior arrangements of storage and marketing and avoids losses. Many researchers have been conducted to develop an efficient method for yield prediction but focus have been always on statistical techniques and not much has been done in machine learning approach.

The crop production depends on various factors which change with every square meter and depends on:

1. Geography of region
2. Weather (Temprature,humidity,percipitation),
3. Soil type (saline, alkaline,sodic,non-alkaline)
4. .Soil composition (ph,N,P,K,EC,OC,Zn,MN,Cu,Fe).

Different subsets of these parameters are used in different prediction models for different crops. Prediction models are basically of two main types:

1. statistics model (e.g. multiple linear regression model) this model uses a single predictive function holding entire sample space
2. Machine learning technique which is emerging technology for knowledge mining that relates input and output variables model

Machine learning: Machine Learning is field of computer science which enables computers to learn without programming, Machine Learning aims at the study and development of algorithms that learn from the data patterns and then can be used to make predictions on new set of data. These algorithms don’t follow static program instructions and they make use of data driven predictions and decisions. Thus in Machine learning approaches the whole emphasis is on data more accurate the data better are the predictions and decisions. Machine Learning is a branch of AI which aims at formulating computational methods for accumulating, changing and updating knowledge in the intelligent systems. Machine learning is act of training computers to optimize a performance criterion using example data or past experience. The model is defined for some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. Machine learning models may be:

1. Predictive
2. Descriptive

A major focus of machine learning is the design of algorithms that recognize complex patterns and make intelligent decisions based on input data. Many machine learning systems aim at eliminating the need of human intervention in the analysis of data, while as others adopt a collaborative approach between machines and human

Machine learning approaches can be of three basic categories:

: a) Supervised learning: This type of learning is also called learning with a teacher, in this a function is generated that maps the inputs to the desired outputs which are pre-classified into different classes

b) Unsupervised learning: This type of learning is also called learning without a teacher, in this the classification of the output into different classes is not already done and the learning algorithm does this classification

c) Reinforcement learning: This type of learning is also called learning with a critic; this involves learning how to act in a given situation when some observations are given as an input.

This paper examines the application of machine learning approaches in prediction of rice yield in Jammu region. The dataset collected from Soil Testing Laboratory Krishi Bhawan Jammu consists the soil composition parameters (Ph,N,P,K,OC,EC,S,Zn,Cu,Mn,Fe).The KNN,Naïve Bayes, and Decision trees have been used .

K-Nearest Neighbor [1] does not have any learning phase, because every time a classification is performed it uses a training set. The assumption behind the k-nearest neighbor algorithm is that a similar classification is produced by similar samples. The similar known samples used for assigning a classification to an unknown sample are described by the parameter.

Naive Bayes [2] classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting.

Decision tree classifier [3] splits entire sample space recursively into smaller sub-sample space which is enough to be formulated by a simple model. The root node (first node) in the tree holds entire sample space. Splitting sample space into smaller sub-sample space means forking root node into children nodes where each child node may be recursively split into leaf nodes (a node on which further split is not possible). The Nodes except leaf node in the tree, split sample space based on a set of condition(s) of the input attributes values and the leaf node assign an output value for those input attributes which are on the path from root to the leaf in the tree. The ultimate goal of sub-sample using decision tree method is to mitigate mixing of different outputs values and assign single output value for subsamples space. The splitting criteria of a node are an impurity measure (e.g. standard deviation used in ID3 algorithm; Gini-Index used in C4.5 algorithm) and Node size (number of data present on a node).

2. RELATED WORK

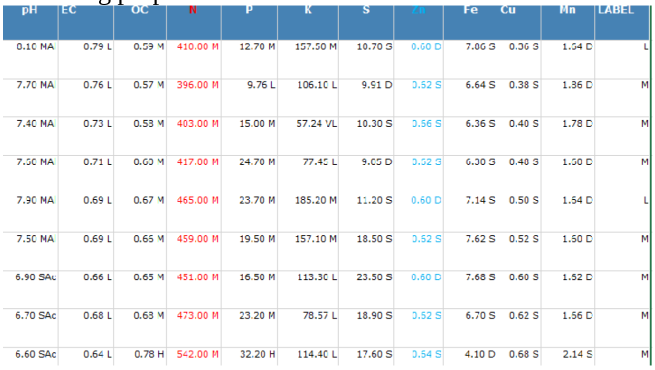
Machine learning in Agriculture is a Novel field still a lot of work has been done in field of Agriculture using Machine learning. [R.Sujatha et.al.[4]] proposed a yield prediction model which used Data Mining techniques for classification and Prediction.This model worked on input parameters crop name, 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. [Shweta Taneja et.al.[5]] 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. [D Ramesh et.al.[6]] proposed a crop yield prediction model that implanted two Data mining techniques namely Multiple Linear Regression and Density Based Clustering techniques. The predict ants were Year', 'Rainfall', 'Area of Sowing', 'Yield', 'Fertilizers' (Nitrogen, Phosphorous and Potassium) and Response variable was 'Production'. In Kg/Hectares. A. Mucherino et.al.[7] conducted a study on the different data mining 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.

3. METHODOLOGY

This section describes the basic steps followed in achieving the goal of prediction of crop yield using Machine Learning approaches. The Proposed Methodology is described in this section. The following steps have been performed to achieve the objectives:

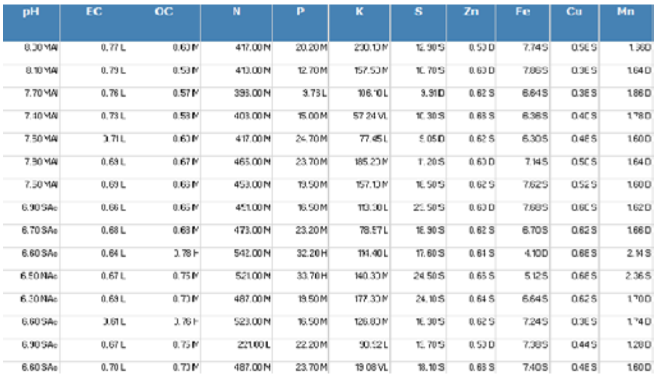
Problem study: A brief study of problems related to maximization of the productivity and prediction of crop yield has been done by going through the related literature review, and with the brief discussions with soil analystsand farmers and broader view of research problem has been gained.

Data collection: After gaining the insight of Research problem the related data has been collected from Soil Testing Laboratory Krishi Bhawan Talab-Tillo Jammu. In hard copy format .The dataset has been collected from villages of RS Pura block and Bishnah block and Marh blocks of Jammu District. The dataset consists of Soil Nutrient status indiduval fieldwise.Each Dataset consists of 1000 samples and total of 10000 samples of data are available. Further data has been divided for training and testing purposes.6062 samples are used for training and 3623 data samples are used for testing purposes, Data has been preprocessed and has been transformed into two excel sheets one for training and one for testing purposes.



pH	EC	OC	N	P	K	S	Zn	Fe	Cu	Mn	LABEL
8.10 MA	0.79 L	0.59 M	410.00 M	12.70 M	157.50 M	10.70 S	0.60 D	7.06 S	0.36 S	1.54 D	L
7.70 MA	0.76 L	0.57 M	396.00 M	9.76 L	106.10 L	9.91 D	0.52 S	6.64 S	0.38 S	1.36 D	M
7.40 MA	0.73 L	0.53 M	403.00 M	15.00 M	57.24 VL	10.30 S	0.56 S	6.36 S	0.40 S	1.78 D	M
7.50 MA	0.71 L	0.60 M	417.00 M	24.70 M	77.45 L	9.05 D	0.52 S	6.30 S	0.40 S	1.50 D	M
7.90 MA	0.69 L	0.67 M	465.00 M	23.70 M	185.20 M	11.20 S	0.60 D	7.14 S	0.50 S	1.54 D	L
7.50 MA	0.69 L	0.65 M	459.00 M	19.50 M	157.10 M	18.50 S	0.52 S	7.62 S	0.52 S	1.50 D	M
6.90 SAc	0.66 L	0.65 M	451.00 M	16.50 M	113.30 L	23.50 S	0.60 D	7.68 S	0.60 S	1.52 D	M
6.70 SAc	0.68 L	0.63 M	473.00 M	23.20 M	78.57 L	18.90 S	0.52 S	6.70 S	0.62 S	1.56 D	M
6.60 SAc	0.64 L	0.78 H	542.00 M	32.20 H	114.40 L	17.60 S	0.54 S	4.10 D	0.68 S	2.14 S	M

Figure1: Training data



pH	EC	OC	N	P	K	S	Zn	Fe	Cu	Mn
8.30 MAl	0.77 L	0.60 M	417.00 N	20.20 M	230.10 M	12.90 S	0.50 D	7.74 S	0.58 S	1.30 D
8.10 MAl	0.79 L	0.53 M	413.00 N	12.70 M	157.50 M	10.70 S	0.60 D	7.86 S	0.38 S	1.64 D
7.70 MAl	0.76 L	0.57 M	398.00 N	3.73 L	106.10 L	3.91 D	0.62 S	6.64 S	0.38 S	1.86 D
7.40 MAl	0.73 L	0.58 M	403.00 N	15.00 M	57.24 VL	10.30 S	0.68 S	6.36 S	0.40 S	1.78 D
7.50 MAl	0.71 L	0.60 M	417.00 N	24.70 M	77.45 L	5.05 D	0.62 S	6.30 S	0.46 S	1.60 D
7.90 MAl	0.69 L	0.67 M	465.00 N	23.70 M	185.20 M	11.20 S	0.60 D	7.14 S	0.50 S	1.64 D
7.50 MAl	0.69 L	0.63 M	453.00 N	19.50 M	157.10 M	16.50 S	0.62 S	7.62 S	0.52 S	1.60 D
6.90 SAc	0.66 L	0.65 M	451.00 N	16.50 M	113.30 L	21.50 S	0.60 D	7.68 S	0.60 S	1.62 D
6.70 SAc	0.68 L	0.68 M	473.00 N	23.20 M	78.57 L	16.90 S	0.62 S	6.70 S	0.62 S	1.66 D
6.60 SAc	0.64 L	0.78 H	542.00 N	32.20 H	111.40 L	17.60 S	0.61 S	4.10 D	0.68 S	2.14 S
6.50 NAc	0.67 L	0.75 M	521.00 N	33.70 H	140.30 M	24.50 S	0.65 S	5.12 S	0.68 S	2.36 S
6.30 NAc	0.69 L	0.73 M	487.00 N	19.50 M	177.30 M	24.10 S	0.64 S	6.64 S	0.62 S	1.70 D
6.60 SAc	0.61 L	0.76 H	523.00 N	16.50 M	126.80 M	16.30 S	0.62 S	7.24 S	0.38 S	1.74 D
6.90 SAc	0.67 L	0.75 M	221.00 L	22.20 M	90.52 L	15.70 S	0.53 D	7.38 S	0.44 S	1.20 D
6.60 SAc	0.70 L	0.73 M	487.00 N	23.70 M	19.08 VL	18.10 S	0.63 S	7.40 S	0.48 S	1.60 D

Figure2: Testing data

Parameters study: The Dataset collected consists of soil composition parameters and is one of subsets for the prediction of yield. The data 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.

Training Data Categorization: The indiduval tuple values of each parameter is classified into LOW,HIGH and MEDIUM category based on critical limits already defined by soil chemists for soils all over India as shown in Table 1

Table 1:Critical Limits of various attributes

Elements	Very low	Low	Medium	High	Very high
pH	<5.0	5.1 - 6.5	6.6 -7.5	7.6 - 8.0	>8.0
Organic carbon(OC) in %	<0.25	0.50- 0	0.51- 0.75	0.76 - 1.00	>1.00
Nitrogen (N) in kg/ha	<150	151 - 250	251 - 400	401 - 600	>600
Phosphorus (P) in kg/ha	<5	6 -10	11 -20	21 - 40	>40
Potassium (K) in kg/ha	<200	201 - 250	251 - 400	401 - 600	>600
Sulphur (S) in kg/ha	<10	11 - 20	21 -30	31 - 40	>40
Zinc (Zn) in mg/kg	<0.30	0.31 -0.60	0.61- 1.20	>1.20	Not Defined
Iron (Fe) in mg/kg	Not Define	<4.5 0	4.51 - 9.0	>9. 0	Not Defined

The LABEL or Response parameters have been calculated on the basis of LAW OF MINIMUIM in Agriculture since in Rice Production in Jammu region depends mainly on two parameters Nitrogen and Zinc Nutrient in the soil and law of minimum holds in it as follows:

NITROGEN(N)	ZINC(Zn)	LABEL
LOW	S	LOW
MEDIUM	S	MEDIUM
HIGH	S	HIGH
LOW	D	LOW
MEDIUM	D	LOW
HIGH	D	LOW

Table 1: Defining Category

Applying Machine Learning Approaches: To achieve the objective the Statistical technique of K fold Cross Validation has been used and three basic machine learning Classifiers are used which are as follows: 1.KNN Classifier 2.Naive Bayes Classifier and Decision Tree Classifiers indepdently on the Data set and Comparative analysis has been done.

A. *K Nearest Neighbour* K-Nearest Neighbor [8] makes predictions based on the outcome of the K neighbors closest to that point. Therefore, to make predictions with KNN, we need to define a metric for measuring the distance between the query point and cases from the examples sample. One of the most popular choices to measure this distance is known as Euclidean (1).

$$D(x,p) = \sqrt{(x-p)2}(\ 1)$$

Where x and p are the query point and a case of the examples sample, respectively. Since KNN predictions are based on the intuitive assumption that objects close in distance are potentially similar, it makes good sense to discriminate between the K nearest neighbors when making predictions. Let the closest points among the K nearest neighbors have more say in affecting the outcome of the query point. This can be achieved by introducing a set of weights W. (2), one for each nearest neighbor, defined by the relative closeness of each neighbor with respect to the query point:

$$\mathrm{W(x,p_i)=exp}(-D(x,pi))/\sum_{i=1}^{k}\mathrm{exp}(-D(x,pi)) \qquad (2)$$

Where D(x, pi) is the distance between the query point x and the ith case pi of the example sample. The weights defined in this manner above will satisfy:

$$\sum_{i=1}^{k}(x0,xi)\ =1 \qquad (3)$$

Thus, for classification problems, the maximum of y is taken for each class variables, as shown:

$$\max y = \sum_{i=1}^{k}(w(x0,xi)yi) \qquad ..\ (4)$$

B. ***Naïve Bayes Classifier****:*

Naive Bayes [9] classifiers can handle an arbitrary number of independent variables, whether continuous or categorical. Given a set of variables, X = {x1, x2, x3..., xd}, we want to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1, c2, c3..., cd}. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable.

Using Bayes' rule:

$$p(Cj|x1,x2,x3,,xd) \propto p(x1,x2,x3\ldots xd|p(Cj)).P(Cj)) \qquad (5)$$

Where p(Cj | x1, x2, x3..., xd) is the posterior probability of class membership, i.e., the probability that X belongs to Cj. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent we can decompose the likelihood of a product of terms:

$$p(X|Cj) \propto \prod_{k=1}^{d} p(x_k|Cj) \qquad (6)$$

And rewrite the posterior as:

$$p(Cj|X) \propto p(Cj)\prod_{k=1}^{d} p(x_k|Cj) \qquad (7)$$

Using Bayes' rule above, we label a new case X with a class level Cj that achieves the highest posterior probability.

C. *Decision Tree Classifier*:

Decision tree learning splits entire sample space recursively into smaller sub-sample space which is enough to be formulated by a simple model [10]. The root node (first node) in the tree holds entire sample space. Splitting sample space into smaller sub-sample space means forking root node into children nodes where each child node may be recursively split into leaf nodes (a node on which further split is not possible). The Nodes except leaf node in the tree, split sample space based on a set of condition(s) of the input attributes values and the leaf node assign an output value for those input attributes which are on the path from root to the leaf in the tree. The ultimate goal of sub-sample using decision tree method is to mitigate mixing of different outputs values and assign single output value for subsamples space. The splitting criteria of a node are an impurity measure (e.g. standard deviation used in ID3 algorithm; Gini-Index used in C4.5 algorithm) and Node size (number of data present on a node). There are many algorithms to build decision tree are: CART [11], M5 [12], and M5-Prime [13]. All these algorithms are similar in tree generation procedure, but they differ in following aspects: first the impurity measure such as M5 uses standard deviation and CART uses variance. Second is prune rule used to avoid over-fitting of a model. Third is the leaf value assignment. M5 apply linear model at leaf nodes instead of constant value Furthermore, M5 is simple, smooth and more accurate than CART algorithm [14]. M5-Prime is subsequent version of M5 dealing with missing values and enumerated attributes.

4. RESULTS

Three Machine learning approaches have been applied indiduval using the Cross Validation techniques with k folds and accuracy of prediction has been observed for each of them. The Rapid Miner Tool version 7.4 has been used in this work. The Training data consists of 11 attributes along with an additional attribute as Label or Response attribute pre-defined by the Soil Testing Lab on the basis of availability of Macro and Micro Nutrients present in the soil. In Rapid Miner Tool the Training Data with 6062 samples have been used to train model separately by KNN, Naïve Bayes and Decision Trees (Simple,ID3,).Once the model has been trained efficiently it is applied on the Testing data set which is different from the Training data in sample values and consists of values from 4 different blocks of Jammu Region other than RS Pura Block. Prediction is done by using the Label attribute in training dataset as Response parameter or Decider Parameter for Prediction.This complete process of Prediction in Rapid Miner is shown in figure 4.

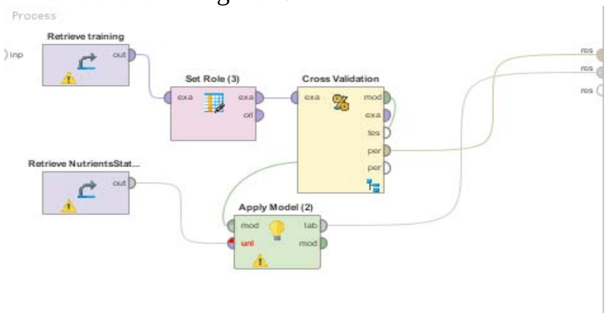


Figure 4: Cross Validation process

The first block is used to Retrieve Training data .The Second block of the Main process is the Cross-validation process block. This block consists of sub processes and each sub process consists of separate Training and Testing processes and in each of this sub process we can use different Classifiers for training process. The second Retrieve block is used to retrieve Testing data from the repository and Apply model operator applies the trained model to the testing data to perform the prediction using the Response Parameter of training Data. Thus training data is input to the Cross Validation process and trained Model and Testing data are input to the Apply Model operator.

The Cross validation subprocess with KNN Classifier is shown in Figure 5.

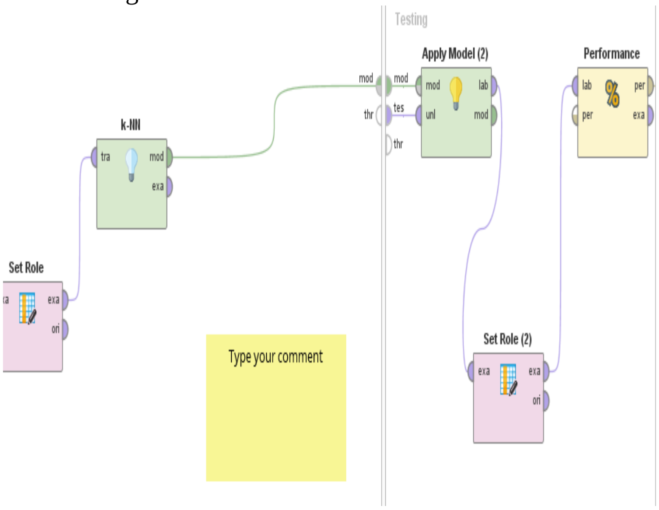


Figure 5:KNN Subprocess

The Training process holds the KNN Classifier which is applied on training data applied from tra port. The Testing subprocess holds the Apply model which applies the Trained Model on the Testing data and Performance operator which measures the accuracy and correctness of the whole process. The same process is repeated for the Naïve Bayes Classifier and Decision Trees.

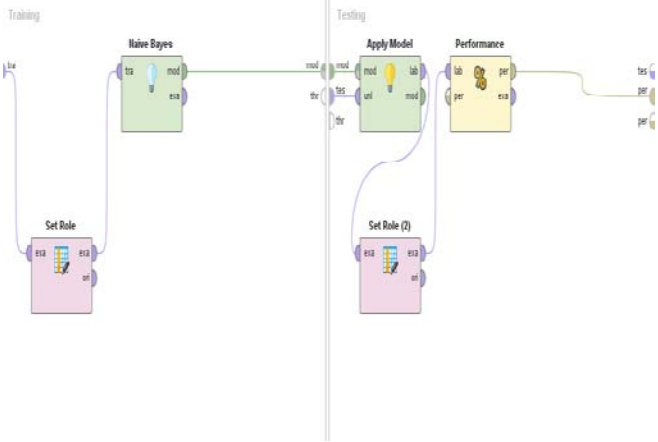


Figure 6:NaïveBayes Subprocess

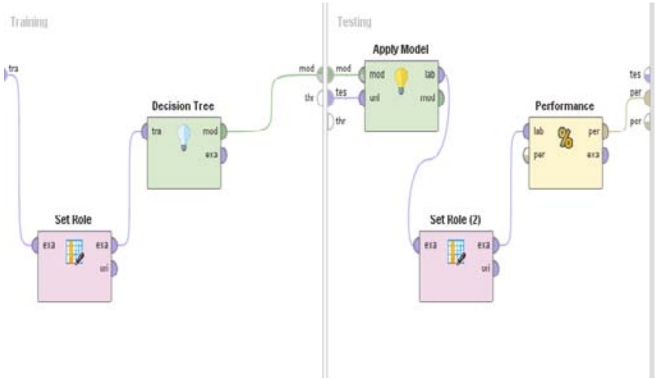
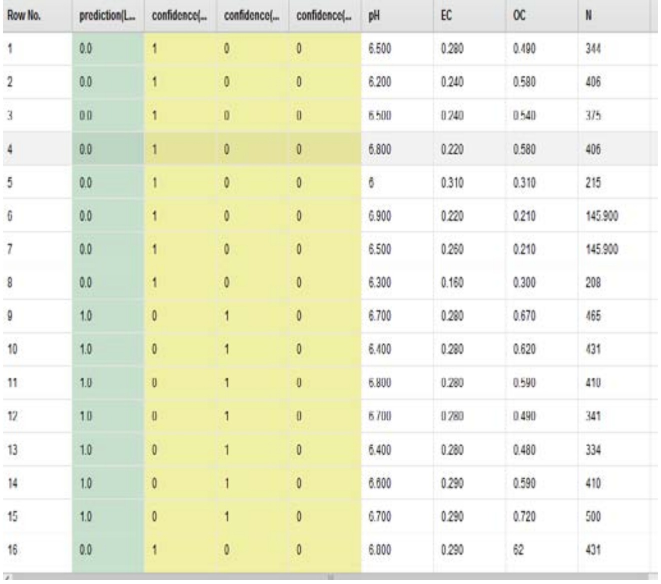


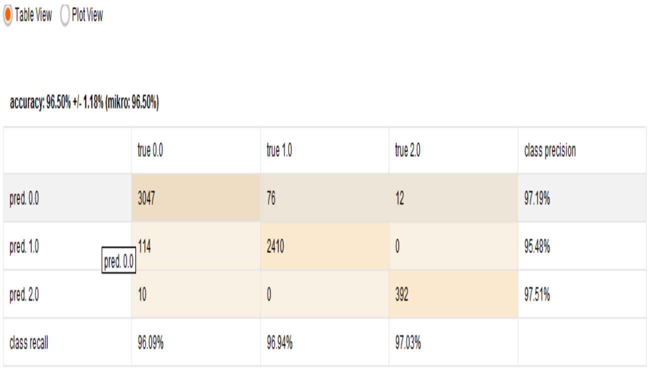
Figure 7: Decision Tree Subprocess

The results of prediction of category of Testing Dataset along with accuracy has been shown in figures below.



Row No.	prediction(L...	confidence(...	confidence(...	confidence(...	pH	EC	OC	N
1	0.0	1	0	0	6.500	0.280	0.490	344
2	0.0	1	0	0	6.200	0.240	0.580	406
3	0.0	1	0	0	6.500	0.240	0.540	375
4	0.0	1	0	0	6.800	0.220	0.580	406
5	0.0	1	0	0	6	0.310	0.310	215
6	0.0	1	0	0	6.900	0.220	0.210	145.900
7	0.0	1	0	0	6.500	0.260	0.210	145.900
8	0.0	1	0	0	6.300	0.160	0.300	208
9	1.0	0	1	0	6.700	0.280	0.670	465
10	1.0	0	1	0	6.400	0.280	0.620	431
11	1.0	0	1	0	6.800	0.280	0.590	410
12	1.0	0	1	0	6.700	0.280	0.490	341
13	1.0	0	1	0	6.400	0.280	0.480	334
14	1.0	0	1	0	6.600	0.290	0.590	410
15	1.0	0	1	0	6.700	0.290	0.720	500
16	0.0	1	0	0	6.800	0.290	62	431

Figure 8: KNN Results



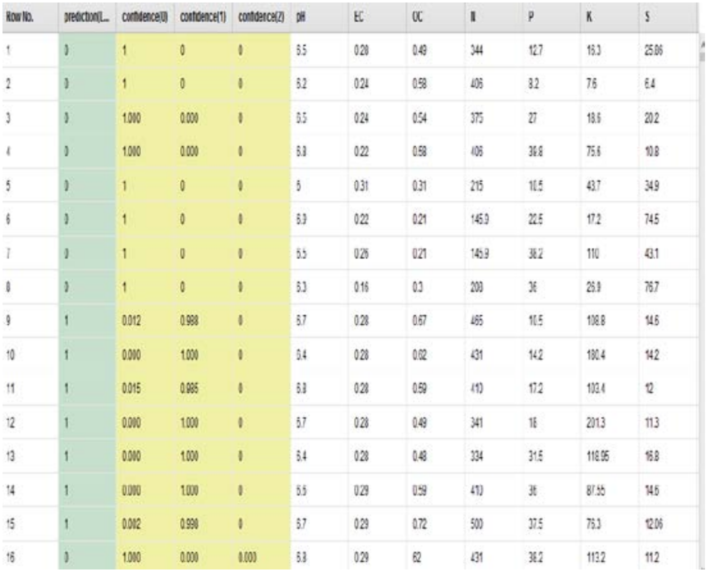
● Table View ○ Plot View

accuracy: 96.50% +/- 1.18% (mikro: 96.50%)

	true 0.0	true 1.0	true 2.0	class precision
pred. 0.0	3047	76	12	97.19%
pred. 1.0	114	2410	0	95.48%
pred. 2.0	10	0	392	97.51%
class recall	96.09%	96.94%	97.03%	

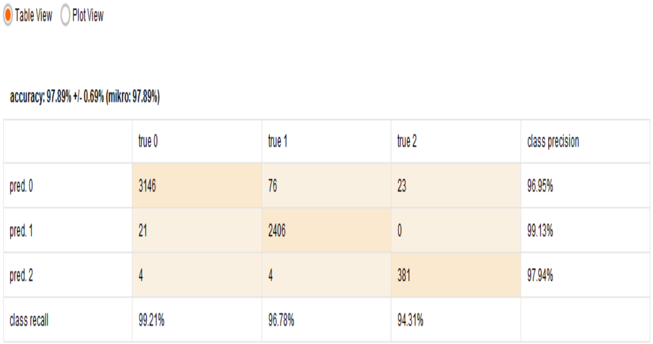
pred. 0.0

Figure 9: Confusion Matrix KNN



Row No.	prediction(L...	confidence(0)	confidence(1)	confidence(2)	pH	EC	OC	N	P	K	S
1	0	1	0	0	6.5	0.28	0.49	344	12.7	16.3	25.06
2	0	1	0	0	6.2	0.24	0.58	406	8.2	7.6	6.4
3	0	1.000	0.000	0	6.5	0.24	0.54	375	27	18.6	20.2
4	0	1.000	0.000	0	6.8	0.22	0.58	406	39.8	75.6	10.8
5	0	1	0	0	6	0.31	0.31	215	10.5	43.7	34.9
6	0	1	0	0	6.9	0.22	0.21	145.9	22.5	17.2	74.5
7	0	1	0	0	6.5	0.26	0.21	145.9	38.2	110	43.1
8	0	1	0	0	6.3	0.16	0.3	208	36	25.9	76.7
9	1	0.012	0.988	0	6.7	0.28	0.67	465	10.5	108.8	14.6
10	1	0.000	1.000	0	6.4	0.28	0.62	431	14.2	180.4	14.2
11	1	0.015	0.985	0	6.8	0.28	0.59	410	17.2	103.4	12
12	1	0.000	1.000	0	6.7	0.28	0.49	341	18	201.3	11.3
13	1	0.000	1.000	0	6.4	0.28	0.48	334	31.5	118.95	16.8
14	1	0.000	1.000	0	6.5	0.29	0.59	410	36	87.55	14.6
15	1	0.002	0.998	0	6.7	0.29	0.72	500	37.5	75.3	12.06
16	0	1.000	0.000	0.000	6.8	0.29	62	431	38.2	113.2	11.2

Figure10:Naïve Bayes Results

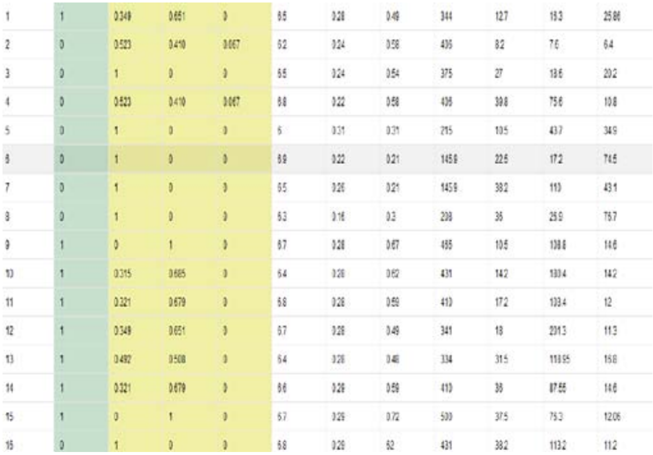


● Table View ○ Plot View

accuracy: 97.89% +/- 0.69% (mikro: 97.89%)

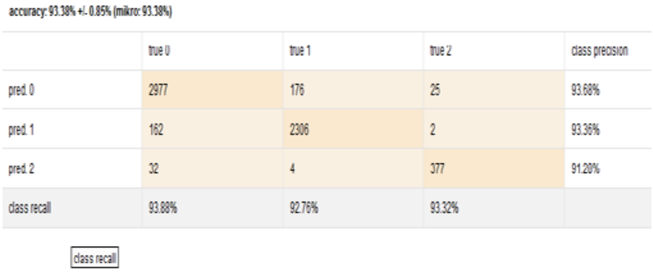
	true 0	true 1	true 2	class precision
pred. 0	3146	76	23	96.95%
pred. 1	21	2406	0	99.13%
pred. 2	4	4	381	97.94%
class recall	99.21%	96.78%	94.31%	

Figure 11:Confusion matrix Naïve Bayes Results



1	1	0.349	0.651	0	6.5	0.28	0.49	344	12.7	16.3	25.86
2	0	0.523	0.410	0.067	6.2	0.24	0.58	406	8.2	7.6	6.4
3	0	1	0	0	6.5	0.24	0.54	375	27	18.6	20.2
4	0	0.523	0.410	0.067	6.8	0.22	0.58	406	39.8	75.6	10.8
5	0	1	0	0	6	0.31	0.31	215	10.5	43.7	34.9
6	0	1	0	0	6.9	0.22	0.21	145.9	22.5	17.2	74.5
7	0	1	0	0	6.5	0.26	0.21	145.9	38.2	110	43.1
8	0	1	0	0	6.3	0.16	0.3	208	36	25.9	76.7
9	1	0	1	0	6.7	0.28	0.67	465	10.5	108.8	14.6
10	1	0.315	0.685	0	6.4	0.28	0.62	431	14.2	180.4	14.2
11	1	0.321	0.679	0	6.8	0.28	0.59	410	17.2	103.4	12
12	1	0.349	0.651	0	6.7	0.28	0.49	341	18	201.3	11.3
13	1	0.492	0.508	0	6.4	0.28	0.48	334	31.5	118.95	15.8
14	1	0.321	0.679	0	6.6	0.29	0.59	410	36	87.55	14.6
15	1	0	1	0	6.7	0.29	0.72	500	37.5	75.3	12.06
16	0	1	0	0	6.8	0.29	62	431	38.2	113.2	11.2

Figure 12:Decision Tree Results



accuracy: 93.38% +/- 0.85% (mikro: 93.38%)

	true 0	true 1	true 2	class precision
pred. 0	2977	176	25	93.68%
pred. 1	162	2306	2	93.36%
pred. 2	32	4	377	91.28%
class recall	93.88%	92.76%	93.32%	

class recall

Figure 13:Confusion matrix Decision Tree Classifier

The Category for Testing Dataset has been predicted as LOW, MEDIUM and HIGH. In each of the processes the confidence values for LOW, MEDIUM and HIGH categories is calculated. These confidence values will help to predict the quality of soil, according to the values of nutrients and micronutrients present in it The Predicted Category of the soil is one having the maximum Confidence value. The Accuracy of the three Classifiers in Prediction are as under:

KNN Classifier=96.43%

Naïve Bayes Classifier=97.80%

Decision Tree Classifier=93.38%

Thus we can see from results that Decision Tree Classifier and Bayes Classifier turn out to be better for classifying the soils into categories and in the prediction of yield on the basis of Nutrient status in the soil.

5. CONCLUSION

In this proposed work the Classification of Soils and prediction of yield has been done on the basis of Soil Composition parameters which is one of subsets for Rice Yield Prediction.This Future work can be proposed by considering other important subsets which are Rain ,Humidity,Temprature Precipitation. Further if indiduval farmer's production data is available it can be used as Response parameter to Predict actual production based on input parameters. Additional Machine Learning Techniques like SVM, Rule Based Induction, can be used.

6. REFERENCES

[1] Cover TM, Hart PE,"K Nearest Neighbor pattern classification", IEEE Trans Info Theory 13(1) : 21-27, 1967.

[2] P.Bhargavi, Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009

[3] Quinlan JR, "Learning with continuous classes". Proc. AI92, 5th Aust. Joint Conf. on Artificial Intelligence (Adams & Sterling, eds.), World Scientific, Singapore, pp: 343-348, 1992.

[4] R Sujatha,Dr P.Issaki,"A Study of Crop Yield Prediction Using Data Mining Techniques" IEEE 2016

[5] Shweta Taneja, Rashmi Arora, Savneet Kaur, "Mining of Soil Data Using Unsupervised Learning Technique", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 7 No.11, 2012.

[6] D Ramesh , B Vishnu Vardhan, "Data mining technique and applications to agriculture yield data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013

[7] Mucherino, P. Papajorgji, P.M. Pardalos, "Data Mining in Agriculture", Springer, 2009

[8] M.Soundarya, R.Balakrishnan," Survey on Classification Techniques in Data mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[9] P.Bhargavi, Dr.S.Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009.

[10] Quinlan JR, "Learning with continuous classes". Proc. AI92, 5th Aust. Joint Conf. on Artificial Intelligence (Adams & Sterling, eds.), World Scientific, Singapore, pp: 343-348, 1992.

[11] Breiman L, Friedman JH, Olshen RA, Stone CJ, "Classification and regression trees". Wadsworth, Belmont, CA, USA, 1984.

[12]] Quinlan JR, "Learning with continuous classes". Proc. AI92, 5th Aust. Joint Conf. on Artificial Intelligence (Adams & Sterling, eds.), World Scientific, Singapore, pp: 343-348, 1992

[13] Wang Y, Witten I, "Inducing model trees for continuous classes". Proc. 9th Eur. Conf. Machine Learning (van Someren M &Widmer G, eds), pp: 128-137, 1997.

[14]] Uysal I, Altay HG, "An overview of regression techniques for knowl- edge discovery". Knowl Eng. Rev 14: 319-340, 1999.