International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V68I4P104

Volume 68 Issue 4, 14-20, April 2020

Original Article

Proactive Crop Supervision with Machine Learning Algorithms for Yield Improvement

Kusum Lata[1], Sajidullah S. Khan[2]

[1]*Research Scholar, Department of SOCSE, Sandip University, Nashik, India*
[2]*Associate Professor, Department of SOCSE, Sandip University, Nashik, India*

Received Date: 26 February 2020
Revised Date: 13 April 2020
Accepted Date: 14 April 2020

Abstract *- Machine learning is the revolutionary approach to solving complex tasks in order to obtain the optimal preferred results. In the internet age, a large amount of data is available to analyze and transform into useful information. This analysis of this data is possible by applying machine learning algorithms to create the relations between different data volumes. Here in this paper, we will discuss the available machine learning algorithms which can be implemented to improve the crop yield prediction with the help of agricultural data sets. This will enable the farmers and governments to get the preferred output which will further boost the Indian economy.*

Keywords *- Machine learning, Crop, SVM, Classification and Regression.*

I. INTRODUCTION

India is mainly an agricultural oriented country. Agriculture gave birth to civilization and plays a precarious role in the global economy. India is an agrarian country, and its economy is highly based on crop production. Therefore, agriculture is the backbone of all business in India. The concept of data mining, artificial intelligence, machine learning, and deep learning is getting so popular in every sphere of technology. Crop yield prediction has been an interesting topic for the authors, advisers and farming-related organizations. The main objective of this research work is to provide a methodology so that it can perform an analysis of crop yield production in an effective manner. CYP depends on the various interrelated factors such as soil, weather and management, and one cannot get the accurate result with the traditional measures, so it will become difficult [1]. To properly understand the agriculture data or environment, one can create new opportunities with the help of big data and high-performance computing techniques. Machine learning is defined as a scientific field that gives machines the ability to learn without being strictly programmed [2].

As I have read several research papers, I found that there are numerous types of models such as Principal component regression, partial least squares, Adaptive forecasts, ARIMA model etc. Now the trend is towards developing a system that is a supervised based model, and it will perform its work as a mixed approach, i.e. classification and Regression approach.

The objective of this paper is to describe various machine-learning algorithms available to predict crop yield production.

This chapter gives a brief description of machine learning, the tasks involved, advantages, disadvantages, models used and also the objective of the proposed work.

II. APPLICATIONS OF MACHINE LEARNING IN AGRICULTURAL CONTEXT

Machine learning is ubiquitously applicable during the whole growing &harvesting cycle of the crops. It initiates with seed, soil management, plant growth management, water feed measurement, and it also comes into a role when robots make a choice to determine the maturity level of plant/crop output with the assistance of computer vision concepts.

The areas where Machine Learning algorithms can be applied at every stage of crop production areas listed below:

- Species management
- Species Breeding
- Species Recognition
- Crop management
- Yield Prediction
- Crop Quality
- Disease Detection
- Weed Detection
- Livestock management
- Livestock Production
- Animal Welfare
- Farmer's Little Helper

Generally, machine learning techniques are used in crop management for crop yield prediction. During the study of the literature review, it is found the most popular models in agriculture are Artificial and Deep Neural Networks (ANNs and DL) and Support Vector Machines (SVMs). But in this paper, I will discuss all the available techniques that can be applied to agricultural data sets to generate the output to help the farmers. At present, machine learning solutions attempt to apply to small data

sets, but in future, they can apply to large data sets to illustrate more accurate results.

III. MACHINE LEARNING ALGORITHMS

Machine learning is the field of computer science where new developments have evolved in recent times. Machine learning is an artificial intelligence (AI) stream to grasp new situations via analysis, straining, observation and experience. Machine learning encounters the wide advancement in the field of IT through coverage to the wide-ranging classification of machine learning algorithms, which are described as below:

A. Parametric

The parametric machine learning algorithms simply the mapping to known functional form.

The algorithms involve two steps:

1. Select a form for the function.
2. To study coefficients from the training data.

$$B0+B1\times X1 +B2\times X2-0$$

Here, B0, B1, andB2 are the coefficients of the line that control the intercept and slope. X1 andX2 are two input variables. x, y is the input and output values.

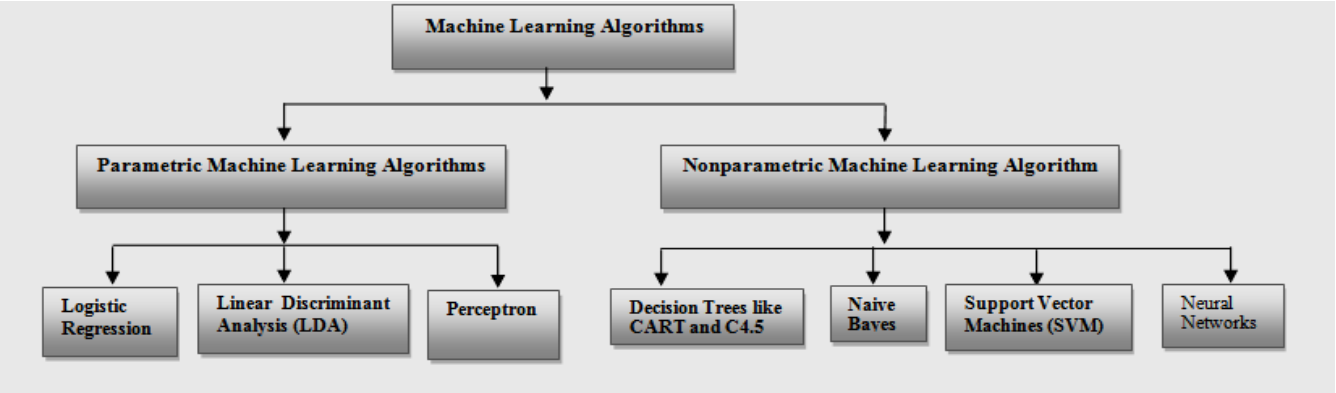

Fig.1 Hierarchal Representation of Machine Learning Algorithms

Benefits of Parametric Machine Learning Algorithms:

- Simpler: These methods are easier to understand and read results.
- Speed: This model is very fast to learn from data.
- Less Data: Less data is required here. This can work well even if the data required is not perfect.

Further detailed classifications of machine learning algorithms are described in the below section:

B. Linear Regression

Linear Regression was developed in the field of statistics, and this can be studied as the relationship between the input and output numerical variables. This is both a statistical and machine learning algorithm. When inputs are single, then this is simple linear Regression, and when inputs are multiple, then this is multiple linear Regression. To prepare and train the linear regression equations to form data, one can make use of different techniques.

The input values (x) and the output value both are numeric.

*The form of model would be: y-B0+B1*x*

B0 and B1 were used for coefficients

C. Logistic Regression

This method is used for binary classification problems, i.e. problems related to two class values. This contains two input variables, i.e. real-valued random variables like (x1, x2) and one output variable(y). This is also called a logistic function.

$$\frac{1}{1+\mathrm{e}^{-\mathrm{value}}}$$

e denotes the base of natural logarithms.

D. Representation Used for Logistic Regression

a) Limitations of Logistic Regression

The limitation of Logistic Regression is two-class problems as it is intended for two classes and binary classification problems. The logistic Regression is rarely used for multiclass, so one can extend it to the multiclass classification. Logistic Regression becomes unstable when the classes are well separated and when it contains few examples from which to estimate the parameters.

b) Linear Discriminant Analysis (LDA)

Linear discriminant analysis is a linear classification technique used when there are more than two classes. It is recommended to try both Logistic Regression and linear discriminant analysis even with binary classification problems.

$$mean_k = \frac{1}{n_k} * \sum_{i-1}^{n} x_i$$

E. Nonparametric Machine Learning Algorithms

Nonparametric algorithms can learn any mapping from inputs to outputs. Benefits of Nonparametric Machine Learning Algorithms:

- Flexibility: This will fit in an enormous number of useful functional forms appropriately.
- Performance: This will give superior performance models for prediction.

a) Limitations of Nonparametric Machine Learning Algorithms

- Additional data required: Training data is essential to determine the mapping function.
- Slower: It is very much slower to train because it contains more parameters.
- Overfitting: Lot of risks to overfitting the training data, and it is not easy to explain the reason behind the complexity of certain parameters

1) Classification and Regression trees

The classification and Regression can also be named CART, which is named by Leo Breiman for the predictive modelling problems. The algorithms such as bagged decision trees boosted decision trees, and the random forest is introduced by CART algorithms.

2) Representation of CART algorithm

The CART model is represented in binary tree form. This model consists of nodes, and where each node is a single input variable, i.e. x and that variable consists of a split point. The variable used here should be numeric. The leaf node contains an output variable, i.e. y, which is used to make a prediction.

Learn a CART model from data:

For Regression predictive modelling problems:

The cost function is the sum squared error across all training samples:

$$\sum_{i=1}^{n}(y_i - prediction_i)$$

Here, y defines the output of the training sample. Prediction is the predicted output.

For classification problems: the Gini cost function is calculated here as this will tell how pure the leaf nodes are.

$$G = \sum_{k=1}^{n} p_k * (1 - p_k)$$

G= Gini cost for overall classes.$p_k = numberoftraininginstances$

F. Bayes Theorem

Bayes theorem is often used in machine learning to select the best hypothesis from the given data. One can make here the use of prior knowledge to select the most probable hypothesis from the given data.

To calculate the probability of a hypothesis can be defined as:

$$P(h/d) = \frac{\mathrm{P}\,(d/h) * P(h)}{P(d)}$$

$P(h/d)$denotes the probability of hypothesis h to the specified data d

P (d/h) denotes the probability of specified data d to the hypothesis h

) $P(h)denotestheprobabilityofhypothesis.$

$P(d)denotestheprobabilityofdata.$

G. Naive Bayes

Naive Bayes is an easy to build, powerful and supervised machine learning algorithm used for predictive modelling by making the use of Bayes theorem**.** Naive Bayes is a classification algorithm that is used to make assumptions for each input variable. For a large number of problem domains, this algorithm plays a vital role.

a) Description used by Naive Bayes

This includes:

- Class Probabilities define the probabilities of each class in the training dataset.
- Conditional Probabilities define the conditional probabilities of each input value given each class value.

b) Gaussian Naive Bayes

The naive Bayes can be extended to real-valued attributes by assuming Gaussian distribution. This expansion of naive Bayes is called Gaussian Naive Bayes. There are other naive Bayes algorithms such as Multinomial naive Bayes and Bernoulli naive Bayes. One can make use of Gaussian naive Bayes to estimate the distribution of data. It is easy to work with this algorithm as it consists of the mean and standard deviation from training data. A Gaussian Naive Bayes algorithm is a unique type of Naive Bayes algorithm. It's particularly used when the features have continuous values and assume that all the features will follow the Gaussian distribution, i.e. normal distribution.

c) Representation of Gaussian Naive Bayes algorithm

One can calculate the mean and standard deviation values for each input variable.

$$mean(x) = \frac{1}{n} * \sum_{i=1}^{n} x_i$$

Here, n denotes the number of instances,x denotes the input variable values from the training data.

$$standarddeviation(x) = \sqrt{\frac{1}{n} * \sum_{i=1}^{n}(x_i - mean(x))^2}$$

Here, n denotes the number of instances,x_iDenotes value of x variable for i[th] instance.

H. K-Nearest Neighbor (KNN)

The k-nearest neighbour algorithm is easy and very effective. This is an easy-to-implement supervised machine learning algorithm that can be used to resolve both classification and regression problems. It doesn't require any kind of training and optimization. KNN is called a sample-based learning technique, where it holds all past data sample space while predicting the target value for the new input sample predictor. It applies distance functions such as Euclidean, Manhattan, Minkowski distance functions to compute the distance from the new

input sample predictor to all training sample predictors and then calculate the k nearest distances with the target values [4]. KNN make use of all data samples during the prediction of new data.

$$Euclideandistance(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

There is some other popular distance measure such as:

- Hamming distance: This calculates the distance between binary vectors.
- Manhattan distance: This calculates the distance between real vectors.
- Minkowski distance: This consists of the generalization of the Euclidean and Manhattan distance.

One can make use of the best distance metric based on the properties of data. If the input variables are similar in type, i.e. measured height and width, the Euclidean is the good distance measure, and if the input variables are not similar in type, i.e. age, gender, height etc., the Manhattan is the good distance measure.

I. Support Vector Machines (SVM)

SVM is one of the most popular algorithms in machine learning. Support

Vectors are simply the coordinates of individual observation. They can use many features without requiring too much computation. Here, one can plot the data points 'n' in dimensional space where n: number of features. Then differentiate it into two classes with the help of classification.

a) Support vector machines (Kernels)

- Linear Kernel SVM
- Polynomial Kernel SVM
- Radial Kernel

IV. COMPARISON OF VARIOUS MACHINE LEARNING ALGORITHMS

In this section, comparison of machine learning algorithms in terms of strength and weakness are discussed in detail and illustrated in below mentioned Table 1:

Table 1. Machine learning algorithms

S No.	Algorithm	Strength	Infirmity
1.	Linear Regression	▪ Linear Regression is straightforward to understand and explain ▪ This can be regularized to avoid overfitting. ▪ The stochastic gradient descent can be updated easily with linear models.	▪ The performance of linear Regression does not hold when they have non-linear relationships. ▪ They are difficult to capture more complex patterns ▪ It takes a lot of time & techniques to add the right interaction terms or polynomials
2.	K-Nearest Neighbor	▪ Performs well on applications that have a sample with many class labels. ▪ This classifier is robust to noisy training data. ▪ A classifier is efficient when the training data is not small. ▪ Versatile—useful for classification or Regression. ▪ High accuracy	▪ Slower than other classification examples ▪ This allocates equal weight to each attribute. ▪ In case there are many irrelevant attributes in the data, it creates ambiguous results. ▪ Results into poor accuracy. ▪ High memory requirement
3.	Decision tree	▪ The decision tree has excellent speed of learning and speed of classification. ▪ Supports transparency of knowledge/classification. ▪ Supports multi-classification.	▪ Even Small variations in the data can show very different looking trees. ▪ Decision tree Construction may affect badly for irrelevant attributes
4.	Naive Bayes	▪ Simple model ▪ Fast ▪ Scalable ▪ Requires little data	▪ Assumes feature independence ▪ Must choose the likelihood function.

5.	K Means	▪ K-Means is the most popular clustering algorithm because it's fast, simple, and flexible if you pre-process your data and engineer useful features.	▪ The number of clusters must be specified ▪ In case the true underlying clusters in the data are not globular, then K-Means produces poor clusters.
6.	Affinity Propagation	▪ No need to give the number of clusters. ▪ But need to specify the sample preference and hyper damping parameters	▪ Quite slow and memory heavy making it difficult to scale to a large number of data sets. ▪ It assumes the true underlying clusters are globular.
7.	Hierarchical /Agglomerative	▪ The main advantage of hierarchical clustering is that the clusters are not assumed to be globular. ▪ It scales well to larger data sets.	▪ Just as in K-Means, the user requires to choose the number of clusters.
8.	DBSCAN	▪ DBSCAN does not assume globular clusters, and its performance is scalable.	▪ DBSCAN is relatively sensitive to hyperparameters such as epsilon and minimum samples.
9.	Support Vector Machine (SVM)	▪ SVM's can model non-linear decision boundaries, and there are many kernels to choose from. They are also comparatively robust against overfitting, especially in high-dimensional space.	▪ However, SVM's are memory intensive, trickier to tune due to the importance of picking the right kernel, and don't scale well to larger data sets.
10.	Classification Tree Ensembles	▪ Perform very well in practice. ▪ They are robust to outliers and scalable. ▪ Able to naturally model the non-linear decision boundaries.	▪ Unconstrained, individual trees are prone to overfitting.
11.	Logistic Regression	▪ This can be updated easily with the new data using stochastic gradient descent. ▪ The outputs have a nice probabilistic interpretation.	▪ Logistic Regression tends to underperform when there are multiple or non-linear decision boundaries. ▪ They are not flexible to naturally capture the more complex relationship.

V. SYSTEM MODEL

Crop yield prediction at the start of crop planning is very important for farmers and governments. In our country, we still use the conservative practice of data collection for crop management and annual yield prediction based on opinion, physical visits and reports. These techniques are one-sided and time intense. The major glitch in the existing crop yield prediction method is accuracy and time-consuming problem. Various studies have been made in crop yield prediction using data mining, but the desired output is not achievable. After analysis of this issue, we are planning to develop some models which will encounter the issues listed below:

- Decide the most favourable planting date.
- Decide the best choice of cultivars.
- Evaluate weather risk.
- Investment decisions
- Policy Decision for Governments.
- Crop Output Time

I have tried to develop a system model for the above problems to further analyze and advise some improvements. The model will be as illustrated below in Fig 2:

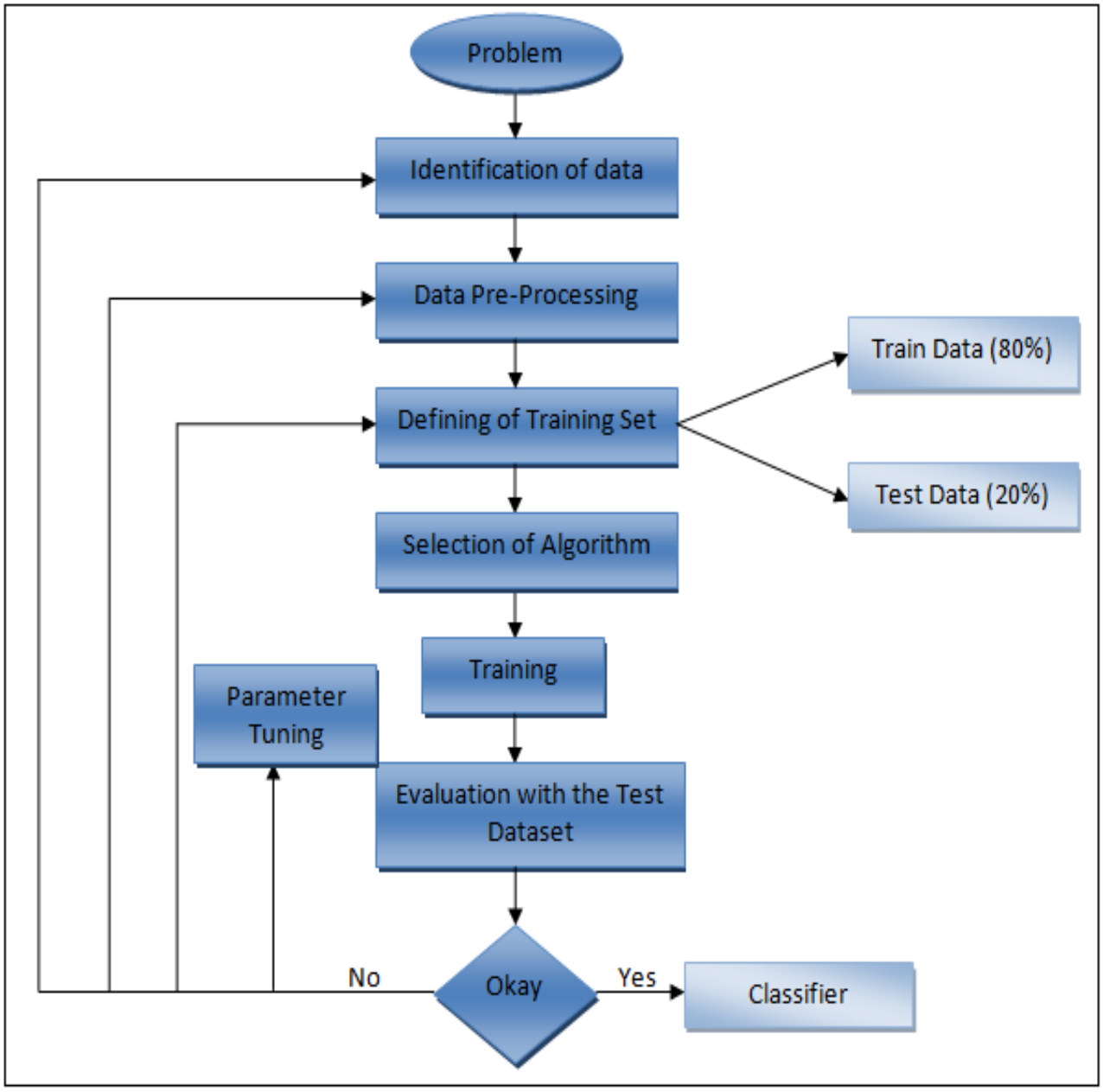

Fig. 2 System Model

In our research, I'll propose a method that would help and suggest the most suitable crop for a specific land-based on analysis of data of previous years on certain affecting parameters by using machine learning algorithms. We are planning to use the previous year's dataset for the analysis, and we'll split that data into two sets, such as training and testing sets. Here, 80% of data will be for training purposes, and 20% of data will be for testing purposes. Our model consists of the entire above mentioned machine learning algorithms, which will guide us to propose a comparison between all of them based on performance.

The current study suggests a machine learning algorithm that, when fed with training data, produces a classifier. The basic machine learning situation for this classifier, we can test with some independent test dataset. We can put this into the classifier and conclude the estimated results. It's really important that training data must be different from testing data. Both training and testing sets are produced by independent sampling from an infinite population.

1. Gathering data from various sources.
2. Cleaning data to have homogeneity.
3. Divide the dataset into two parts:
    - Training (80%)
    - Testing (20%)
4. Model Building: Selecting the right machine learning algorithms.
5. Gaining insights from the model's outcomes.
6. Data visualization: Transforming results into visual graphs.

VI. CONCLUSION – MACHINE LEARNING VS STATISTICS

Agriculture is the most important application area, particularly in developing countries like India. The use of machine learning algorithms in agriculture can transform the situation of decision making, and farmers can yield in a better way. Machine learning plays a very important role in decision making on several issues related to the agriculture field. This paper displays the survey of machine learning techniques for crop yield prediction. This paper integrates the efforts made by various authors in one place, so it is helpful for researchers to obtain information on the current state of machine learning techniques and applications in the context of to agriculture field.

REFERENCES

[1] J. Liu, C. E. Goering, Lei Tian, 2001. A Neural Network For Setting Target Corn Yields. Transactions of The American Society of Agricultural Engineers, 44(3) (2001) 705-713.

[2] Samuela.L, Some Studies in Machine Learning Using the Game of Checkers. Ibm J.Res. Dev., 44 (1959) 206-226.

[3] Marcello Donatelli, Amit Kumar Srivastava, Gregory Duveiller, Stefan Niemeyer and Davide Fumagalli, Climate Change Impact And Potential Adaptation Strategies Under Alternate Realizations of Climate Scenarios for Three Major Crops in Europe, Environmental Research Letters, 10(7) (2015) 075005.

[4] Rakesh Kumar, M.P. Singh, Prabhat Kumar, J.P. Singh, 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), (2015).

[5] Report on Economic Survey of Maharashtra 2012-2013, Directorate of Economics And Statistics, Planning Department, Government of Maharashtra, Mumbai (2013).

[6] D. Diepeveen And L. Armstrong, Identifying Key Crop Performance Traits using Data Mining World Conference on Agriculture, Information And It, (2008).

[7] Alexander Murynin, Konstantin Gorokhovskiy And Vladimir Ignatie, Efficiency of Crop Yield Forecasting Depending on The Moment of Prediction Based on Large Remote Sensing Data Set Retrieved Fromhttp://Worldcomp-Proceedings.Com/Proc/P2013/Dmi8036.Pdf

[8] Hemageetha, N., A Survey on the Application of Data Mining Techniques to Analyze the Soil for Agricultural Purpose, 3rd International Conference on Computing for Sustainable Global Development (Indiacom), (2016) 3112-3117.

[9] Wu Fan, Chenchong, Guoxiaoling, Yu Hua, Wang Juyun. Prediction of Crop Yield using Big Data. 8th International Symposium on Computational Intelligence and Design (Iscid),1 (2015) 255-260.

[10] Monali Paul, Santosh K. Vishwakarma, Ashok Verma. Analysis Of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach. Computational Intelligence and Communication Networks (Cicn), (2015) 766-771.

[11] Subhadra Mishra, Debahuti Mishra, Gourharisantra, Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper, Indian Journal of Science and Technology, 9(38) (2016) 1-14

[12] Kushwaha, A.K., Swetabhattachrya, Crop Yield Prediction using Agro Algorithm in Hadoop, International Journal of Computer Science and Information Technology & Security (Ijcsits), 5(2) (2015) 271-274.

[13] Sujatha, R., Isakki, P., A Study on Crop Yield Forecasting Using Classification Techniques, International Conference on Computing Technologies and Intelligent Data Engineering (Icctide), (2016) 1-4.

[14] N.Gandhi And L.J. Armstrong, Applying Data Mining Techniques To Predict The Yield of Rice In Humid Subtropical Climatic Zone Of India, Proceedings of The 10th Indiacom-2016, 3rd 2016 Ieee International Conference on Computing for Sustainable Global Development, New Delhi, India, 16th To 18th March (2016).

[15] N. Gandhi And L. Armstrong, Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques, Ieee International Conference on Advances in Computer Applications (Icaca), (2016) 357-363.

[16] Shweta Srivastava, Diwakar Yagysen, Implementaion Of Genetic Algorithm for Agriculture System, International Journal of New Innovations in Engineering and Technology, 5(1) (2016).

[17] Shruti Mishra, Priyanka Paygude, Snehal Chaudhary, Sonali Idate, Use of Data Mining in Crop Yield Prediction, 2nd International Conference on Inventive Systems and Control (Icisc), (2018).

[18] Rossana Mc, L. D, A Prediction Model Framework For Crop Yield Prediction. Asia Pacific Industrial Engineering and Management Society Conference Proceedings Cebu, Philippines, (2013)185.

[19] R.Kalpana, N.Shanti And S.Arumugam, A Survey on Data Mining Techniques in Agriculture, International Journal of Advances in Computer Science And Technology, 3(8) (2014) 426- 431.

[20] Dr Shirin Bhanu Koduri, Loshma Gunisetti, Ch Raja Ramesh, K V Mutyalu and D. Ganesh, Prediction of Crop Production Using Adaboost Regression Method, International Conference on Computer Vision and Machine Learning, Conf. Series 1228, (2019).