Survey Paper on Agricultural Dataset for Improving Crop Yield Prediction using Machine Learning Algorithms

Atul Tripathi
MTech 3rd sem (CSE Student)
BUIT, Bhopal

Bhawani Singh Rathore
Assistant professor (CSE)
BUIT, Bhopal

Divakar Singh
HOD (CSE)
BUIT, Bhopal

ABSTRACT
By being able to anticipate crop yields more precisely than what has already been done in the field, we hope to improve precision agriculture. A system that examines a data set that includes crop, cost of cultivation, cost of production, and crop yield from the past, builds a statistical model through learning, and then tries to assist farmers in making accurate and precise decisions about which crops can be grown profitably in the near future can be created with the aid of machine learning techniques and the appropriate optimizations and fine-tuning of the classifying algorithm.As a consequence of this work, the system would have a set of guidelines (referred to as a "Knowledge base," which learns through additional training and data sets) that aid farmers in selecting the crops that are most likely to yield a profit in the current environment. The method used to categories agricultural datasets by crop, area (Quantel/hectar), and production. Here, we investigate our classification algorithms with the aid of the WEKA tool. There is currently a push to transform the vast amounts of agricultural data into various technologies and make them accessible to farmers so they can make better decisions. This survey study investigates the best machine learning algorithms, including Random Tree, J48, Bayes Net, and KStar. We research machine learning methods to uncover relevant data in the agricultural dataset so that we may more accurately forecast crop yields for important crops.

General Terms
The objective of this study is to survey the available agricultural datasets and evaluate their potential for improving crop yield prediction using machine learning algorithms.

Keywords
WEKA tool, J48, Bayes Net, KStar and Random Tree, Machine learning and Crop Yield Prediction Agriculture, Supervised Algorithms.

1. INTRODUCTION
India's population is expanding swiftly in comparison to the rest of the world. The need for food will increase quickly because there are only so many resources and acres of land available. Approximately 70% of Indians, according to government data, reside in villages. Farmers in communities depend on farming for a living, and the success of their harvests determines how big they may grow. However, there are still a number of issues in the sector of agriculture that need to be resolved. More over half of India's workforce is employed in the sector, which contributes roughly 17–18% of the nation's GDP (GDP). A subfield of computer science called artificial intelligence seeks to emulate human intelligence in machines One of the subfields of artificial intelligence is machine learning. Contrary to other programming applications, machine learning does not require us to explicitly state the procedures or prerequisites[1]. The major objective of this research is to develop a technique for accurately analyzing crop yield output [2]. Determining how much a crop will produce is the main focus of study in this subject. Early predictions of crop yields can aid farmers in avoiding losses and maximizing the output of their crops. We can claim with some certainty that this is a problem that dates back to agricultural planning [8]. Here, our major objective is to test various machine learning algorithms and determine which one provides the most accurate crop yield results. Crop yield estimates are frequently based on the weather, soil, and location [15]. In order to profit from their agricultural output, we also aim to provide farmers with better technologies. Concretely speaking, there are two types of machine learning algorithms: supervised and unsupervised. Both algorithms use raw data to forecast or identify various data trends in order to estimate crop production as precisely as possible. Planning for agriculture is important for a country with an agricultural economy and for ensuring food security [4]. One could predict the crop in the past using the farmer's prior experience with that crop. Data mining is a technique for extracting, structuring, loading, and forecasting relevant information from massive amounts of data in order to identify trends and organize the information for future use [15]. Agriculture is an important sector of the global economy, and crop yield prediction is crucial for farmers and agricultural businesses to make informed decisions about planting, resource allocation, and risk management. Machine learning algorithms have the potential to significantly improve crop yield prediction by analyzing large amounts of data and making more accurate predictions than traditional methods..

2. LITERATURE SURVEY
studies in this part. Diepeveen and Armstrong [6] discuss some of the most crucial elements that influence how well a crop performs in various locations, including the season, the method of planting, and the kind of soil. This study primarily discusses how to use various data mining approaches to assess the performance of crops. Monali Paul and others [10] The crops are analysed and divided into groups in order to forecast how much they will yield. These groups are created using data mining algorithms. The Naive Bayes and K-Nearest Neighbor grouping rules, among others, are discussed in this study. Murynin et al. [7] examine how the forecast's accuracy is impacted by the prediction. Crop simulation is driven by weather data[3]. A linear model is employed to determine the yield. By including non-linear features to this model, the prediction's accuracy is increased. Based on how long it took from the time the forecast was generated until the time of harvest, the model was believed to be reliable. Hemageetha [8] focuses on how to forecast crop yields by taking into account various soil characteristics, like pH, nitrogen, moisture, etc. The Naive Bayes algorithm, which is 77% effective, is used to classify the soil. To match the soil with the crops that can benefit from it the greatest, the Apriori

algorithm is applied. Additionally, the effectiveness of classifying data using Naive Bayes, J48, and JRIP is contrasted.

The study by AakunuriManjula and Dr. G. Narsimha [12] compares classification techniques including KNN, Bayesian Network, and Decision Tree. Santhosh K. Vishwakarma, Ashok Verma, and Monali Paul [13] The examined soil datasets are predicted using a data mining approach, which is used in the process of predicting crop output. The system's failure to take into account the demand that exists in the agricultural economy is a drawback. In our approach, farmers are advised which crops to grow depending on market prices and demand. Zhang Tng [14] Crop yield estimation utilising classification approaches calculates crop yield and, using data mining techniques, chooses the best crop for production, increasing the value and profit of the farming area. The infeasibility of the ways for meeting demand and for providing advice to farmers is one of this system's drawbacks. Crop prediction model framework was built by Rossana MC [16] and it was found that planting procedures, notably the application of the proper quantity of fertiliser, had a significant impact on corn production rather than climate-related variables. An overview of information mining methods in the agribusiness is provided by R. Kalpana [17]. Agriculture data processing is linked to expand the analytical field. The use of information mining techniques is crucial in the connected fields of agriculture and the environment. With the aid of the available knowledge, some problems in agriculture, such as yield estimation and crop productivity, still need to be resolved. The goal of this survey is to identify appropriate data processing models so that high accuracy and prediction skills can be understood. According to this theory, studying additional techniques and algorithms related to agricultural difficulties can offer a wise path for the development of agriculture. Last but not least, using data processing methods improperly in agriculture may be a more modern approach to problem-solving than the conventional and conventional method.

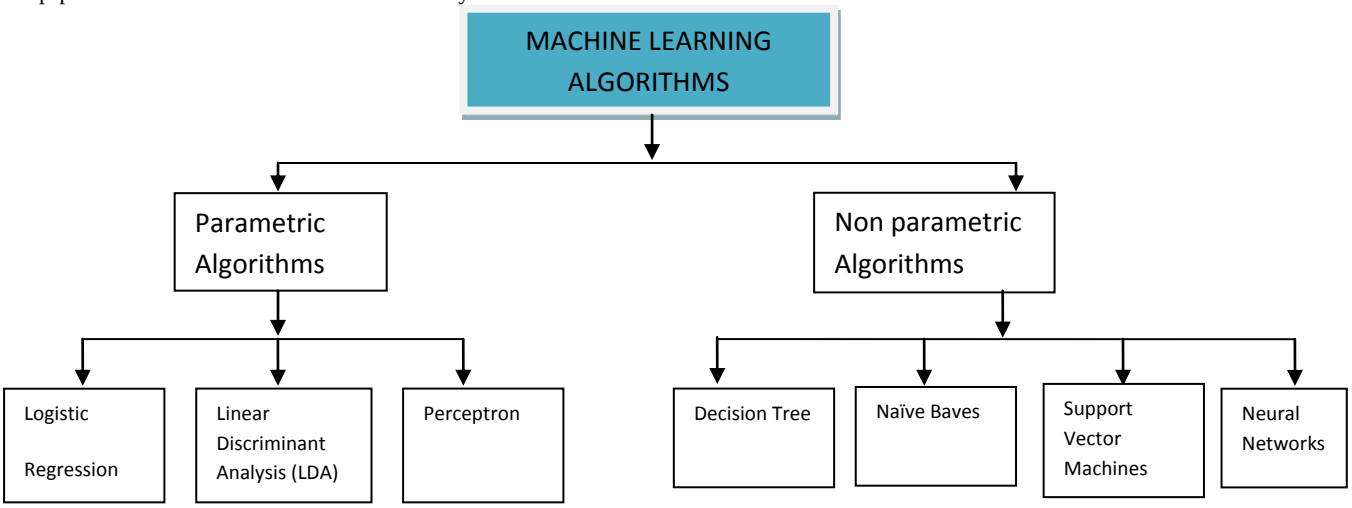

Fig.1 Hierarchal Representation of Machine Learning Algorithms

3. RESEARCH METHODS

The sample study area, agricultural datasets, and firm methodology are the main factors that contribute to this research.

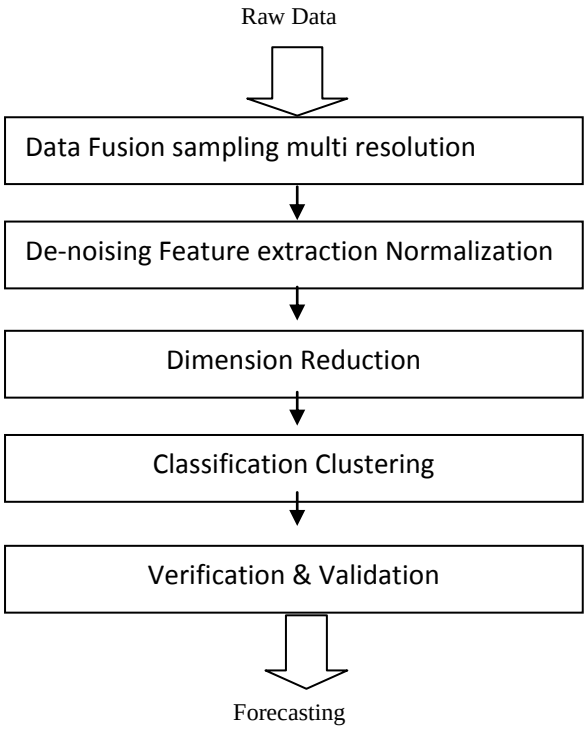

Fig.2 Steps in Knowledge discovery

4. METHODS AND METHODOLOGY

A comprehensive literature review was conducted to identify relevant agricultural datasets. Datasets were selected based on their availability, size, and relevance to crop yield prediction. The following machine learning algorithms were evaluated for their potential to improve crop yield prediction using the selected datasets: linear regression, decision trees, random forests, support vector machines (SVMs), and neural networks.

In this paper, we focus on data preprocessing and use the weka tool to analyse the simulation results that come from using machine learning classifiers. Following Data mining techniques are used to describe a crop yield forecast system by analysing agricultural datasets. [18] Three key variables—total production for each crop, cultivated area, and seasons—are employed in the data set. To analyse the dataset, the WEKA tool is utilised with the J48, LAD Tree, LWL, and IBK classifier algorithms. The Root Mean Squared Error, Mean Absolute Error, and Relative Absolute Error metrics are also used to compare the classifiers.

WEKA Tool: The fuller term for WEKA is "Waikato Environment for Knowledge Analysis." This tool was created at the New Zealand University of Waikato. Several machine learning algorithms have been created in this tool and can be applied immediately to a batch of agricultural data. The Graphical User Interface is a useful tool for contrasting the outcomes of various trials. Artificial intelligence may train itself or learn from data with the aid of machine learning. The Preprocess is Weka. Data sets are preprocessed in a variety of ways in this step.

Classify: A data set that performs classification and regression operations and also evaluates them goes through training and testing.

Cluster: The clustering of dataset takes place.

•**Associate:** Association rules are applied for the dataset and evaluation takes place.

•**Select attributes**: Selection of the most relevant attributes of the given dataset.

•**Visualize:** Visualization of data set occurs in the different two-dimensional plot and interact with them.

The main features of weka are as listed below:

- Platform independent
- Open source and free
- Different machine learning algorithms
- Easy to use
- Data prepossessing tools
- Flexibility for scripting experiments
- Graphical user interface

Weka is a tool that has four keys that are explained below:

Explorer: It is the location where you can explore the data

Experimenter: It is the location where you can execute the experiments and can bear statistical tests between learning outlines.

Knowledge flow: It is the location which upkeep essentially the similar functions as the EXPLORER nevertheless with a drag and drop interface.

Simply CLI: It offers an easy command-line interface that permits implementation of WEKA commands for operating systems and does not permit their own command line interface.

Several agricultural datasets were identified that could potentially be used for improving crop yield prediction. These datasets included weather data, soil data, crop data, and satellite imagery.

Linear regression was found to be a simple and effective algorithm for crop yield prediction, especially when combined with feature selection techniques. Decision trees and random forests performed well in some cases, but were prone to overfitting. SVMs showed good performance in some cases, but were sensitive to the choice of kernel and other hyperparameters. Neural networks showed good performance in most cases, but required more data and computational resources.

5. PERFORMANCE EVALUATION

•In this part, the fundamental evaluation metrics from the confusion matrix are explained. Accuracy, sensitivity, specificity, and precision are four outcomes that are combined to form the confusion matrix. Additionally, this includes the ROC, precision-recall, etc. There are four classification results:

• **True positive (TP):** As the classifier name appears, TP is just the representation of number of correct classifiers reserved to positive class.

•**True negative (TN):** As the classifier name appears,TN is just the representation of number of correct classifiers reserved to negative class.

• **False positive (FP):**As the classifier name appears, FP is just the representation of number of classifiers reserved to positive class but in realism fit in to negative class.

•**False negative (FN):**As the classifier name appears,FN is just the representation of number of classifiers reserved to negative class but in realism fit in to positive class.

Sensitivity (REC/TPR): This is also called as true positive rate and recall. It can be determined as the no. of correct positive predictions divided by the total no. of positives.

Specificity (SP/TNR): This is also called as true negative rate. It can be determined as the no. of correct negative predictions divided by the total no. of negative.

Precision (PPV): This is often referred to as positive forecasting and involves the use of meteorological parameters such as temperature, humidity, wind speed, solar radiation, precipitation, soil temperature, and soil moisture readings from earlier days. On the basis of the historical data, mathematical models are created using the sparse and well-researched machine learning techniques SVM and RVM. The site-specific model, which was created from a Precision Agriculture perspective, may combine data from various sources at the granularity of one day. By enabling the model to interact with human knowledge or trustworthy soil moisture, the feature that accepts user-provided data strengthens the system. Depending on the chosen class, we may see the attributes. The chosen properties can be seen by selecting the "Visualize All" button.

A. Filters

Weka's preprocessing instruments are referred to as filters. One can choose or remove attributes from our dataset, as well as apply filters, depending on our needs. Weka filters are referred to be data preparation tools since they can be used to alter the data sets.

B. Explore: Building classifiers:

In WEKA tool, classifiers are the models used for classification and regression.

• **Choosing a classifier:** Once you have loaded your data set in Weka tool you have to click on the classifier button, and choose theappropriate classifier for your data set and one can startvalue. It can be determined as the no. of correct positive predictions divided by the total no. of positive.

Accuracy: It can be determined as the no. of all correct predictions divided by the

total no. of dataset.

RAE (Relative absolute error): Relative absolute error is just the replica of relative squared error.

RRSE (Relative root mean square error): It is basically applied to calculate the differences between the differentvalues estimated by a model, estimator and the observed values. Sometimes it is also termed as root mean square deviation (RMSD).

MAE (Mean absolute error): The difference between the two continuous variables is called as mean absolute error. For the mean absolute error subtract mean value from each term and calculate the mean of modulus i.e. positive of thosevalues.

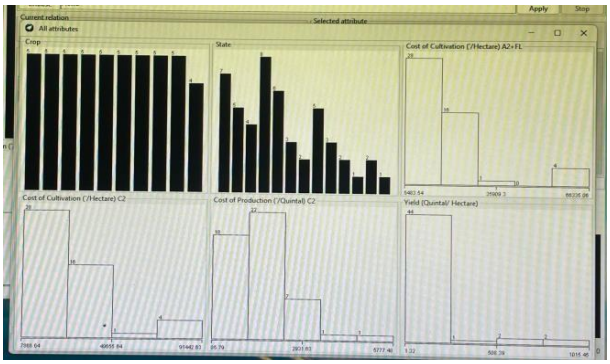

Fig.3 Weka screenshot of crop and all attributes in visualtion form

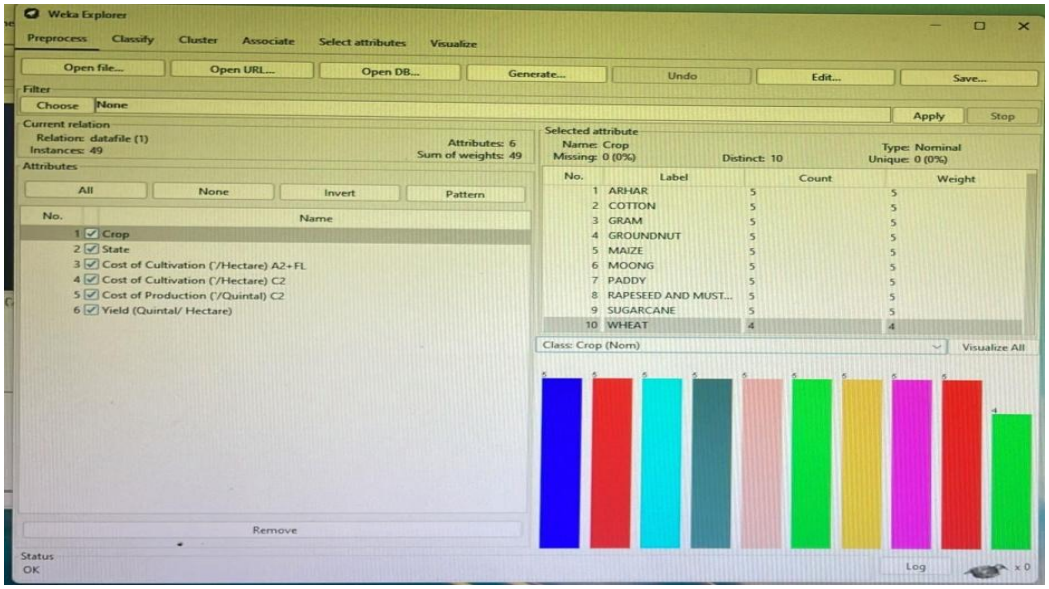

Fig.4 Weka screenshot of classes and their count of selected attributes in visualtion form.

Table I. Study &Comparison of Machine learning algorithms

S No.	Algorithm	Strength	Infirmity
1.	LinearRegression	▪ Linear regression is simple to comprehend and explain. ▪ This can be adjusted to prevent becoming too tight. ▪ With linear models, the stochastic gradient descent can be easily updated.	▪ When there are non-linear relationships, linear regression's effectiveness is not maintained. ▪ It is challenging to capture more intricate patterns with them. ▪ To add the appropriate interaction terms or polynomials, it takes a lot of practice and technical skill.
2.	K-NearestNeighbor	▪ Works well with applications that use samples with lots of class labels. ▪ This classifier is robust to noisy training data. ▪ Classifier is efficient when the training data is not small. ▪ Versatile—useful For classification or regression. ▪ High accuracy	▪ Slightly more slowly than other classification examples ▪ This allocates equal weight to each attribute. ▪ Incasetherearemanyirrelevantattributesintheda ta,itcreatesambiguousresults. ▪ Results into poor accuracy. ▪ High memory requirement
3.	Decisiontree	▪ Decision tree has excellents peed 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.	NaiveBayes	▪ Simple model ▪ Fast ▪ Scalable ▪ Requires little data	▪ Assumes feature independence ▪ Must choose the likelihood function.
5.	KMeans	▪ K-Means is most popular clustering algorithm because it's fast, simple, and flexible if you per-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 damping hyper parameters	▪ Quite slow and memory heavy making it difficult to scale to large amount of datasets. ▪ It assumes the true underlying clusters are globular.
7.	Hierarchical /Agglomerative	▪ Themainadvantageofhierarchical clusteringisthattheclustersarenotassumedtobeglobular. ▪ 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 hyper parameters such as epsilon and minimum samples.
9.	Support Vector Machine (SVM)	▪ SVMs have a wide range of kernel options and can model non-linear decision boundaries. Additionally, they are relatively resistant to over fitting, particularly in high-dimensional space.	▪ SVMs don't scale well to larger data sets, are memory-intensive, and are more difficult to tune because of the significance of choosing the right kernel.
10.	Classification Tree Ensembles	▪ Perform admirably during practice. ▪ They are scalable and resistant to outliers. ▪ Capable of modelling the non-linear decision limits intuitively.	▪ Individual trees are prone to overfitting when unrestrained.
11.	Logistic Regression	▪ With the use of stochastic gradient descent, this may be quickly updated with the fresh data. ▪ The results are nicely interpreted probabilistically.	▪ When there are numerous or non-linear decision boundaries, logistic regression frequently performs poorly. ▪ They lack the adaptability to easily capture

6. CONCLUSION – MACHINE LEARNING VS. STATISTICS

Machine learning algorithms have the potential to significantly improve crop yield prediction by analyzing large amounts of data and making more accurate predictions than traditional methods. Linear regression, decision trees, random forests, SVMs, and neural networks are all potential options for improving crop yield prediction using agricultural datasets. Further research is needed to determine the most effective machine learning algorithms and datasets for specific crops and regions. The recent increase in population calls attention to the need to fulfil the needs of the populace and provide food security and environmental protection. In this regard, crop yield output and early forecasting are quite important. Making better crop planning decisions may be possible with crop yield prediction utilising intelligent machine learning approaches. This proposed study compares a number of machine learning algorithms, including Random Tree, KStar, Bayes Net, and J48, using a dataset to determine which one is the most accurate for predicting crop yields. According to our research, experimental comparisons of various algorithms will yield a range of accuracy findings that may further be helpful to farmers. The main take away from this research is that selecting an appropriate algorithm and multidimensional dataset can provide farmers with early predictions and suggestions for crop productivity. In the future, we will use these algorithms on sizable datasets to propose a seasonal decision support system for Indian farmers that uses weather and non-weather information.

7. RECOMMENDATIONS & FUTURE SCOPE

Future research should focus on identifying the most effective machine learning algorithms and datasets for specific crops and regions. This could involve conducting detailed case

studies or large-scale experiments to evaluate the performance of different algorithms and datasets. It would also be useful to develop new agricultural datasets, particularly those that capture detailed information about crop growth and development. Finally, it would be beneficial to explore ways to integrate machine learning algorithms into existing agricultural systems, such as precision farming tools, to improve decision-making and resource management in the agricultural sector.

8. ACKNOWLEDGMENTS

The authors express their sincere gratitude to the Director Academics and Head of the Department of Computer Science and Engineering of University Institute of Technology Engineering College Barkatullah University for giving constant encouragement and support to complete the work.

9. REFERENCES

[1] 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) 012005.

[2] Kusum Lata,Sajidullah S Khan "Proactive Crop Supervision with Machine Learning Algorithms for Yield Improvement". April 2020 International Journal of Computer Trends and Technology 68(4):14-21

[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, vol. 10, no. 7, Jul 2015, Art. No. 075005.

[4] Rakesh Kumar, M.P. Singh, Prabhat Kumar, J.P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique",2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM),27 August 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" retrievedfromhttp://worldcompproceedings.com/proc/p2013/DMI80 3 6.pdf.

[8] Hemageetha, N., "A survey on application of data mining techniques to analyze the soil for agricultural purpose", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp.3112-3117, 2016.

[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).2015; 1, 255- 260.

[10] Monali Paul, Santosh K. Vishwakarma, Ashok Verma. Analysis of soil behavior 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.2016, 9(38), 1-14

[12] AakunuriManjula, Dr.G .Narsimha (2015), 'XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction', Conference on Intelligent Systems and Control (ISCO)

[13] Monali Paul, Santhosh K. Vishwakarma, Ashok Verma, "Prediction of crop yield using Data Mining Approach" Computational Intelligence and Communication Networks (CICN), International Conference 12-14 Dec. 2015.

[14] Tng Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms", Proceedings of the twenty-first international conference on Machine Learning. Shweta Srivastava, Diwakar Yagysen,"Implementaion of Genetic Algorithm for Agriculture System", International Journal of New Innovations in Engineering and Technology Volume 5 Issue 1 -May 2016.

[15] R.Kalpana, N.Shanti and S.Arumugam, "A survey on data mining techniques in Agriculture", International Journal of advances in Computer Science and Technology, vol. 3, No. 8,426- 431, 2014.

[16] AakunuriManjula, G. Narsimha, "XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction", IEEE Sponsored 9th ISCO, 2015.

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

[18] Shruti Mishra, Priyanka Paygude,Snehal Chaudhary, Sonali Idate, "Use of data mining in crop yield prediction",2018 2nd International Conference on Inventive Systems and Control (ICISC).