# A Survey on Data Mining Approach for Crop Yield Prediction and Soil behavior Analysis

Falguni Prajapati [1], Prof. J. S. Dhobi[2]

[1]Student (Master of Engineering), Computer Engineering, Government Engineering College, Gandhinagar, Gujarat, India

[2]Associate Professor, Computer Department, Government Engineering College, Gandhinagar, Gujarat, India

***Abstract***- **Agriculture is one of the major revenue producing sectors of India and a source of survival. Several seasonal, economic and biological patterns influence the crop production but unpredictable changes in these patterns lead to a great loss to farmers. These risks can be reduced when suitable approaches are employed on data related to soil, temperature, atmospheric pressure, humidity and crop type. Whereas, crop and weather forecasting can be predicted by deriving useful insights from these agricultural data that aids farmers to decide on the crop they would like to plant for the forthcoming year leading to maximum profit. This paper presents a survey on the various algorithms used for crop yield prediction.**

***Index Terms***- **Yield Prediction, Data Mining, Classification Rule, Soil Analysis.**

## I. INTRODUCTION

Agriculture is superior to the human beings, because it forms the basis for food security. Agriculture is the main source of national income for most developing countries [1]. However, for the developed countries, agriculture contributes a larger percentage to their national income. Agriculture is one of the major sectors to be impacted by different sources like climatic changes, soil attributes, seasonal changes etc., [2]. India is predominantly an agriculture based country, and agriculture is the important occupation for most of the Indian families. In India, over 60.3% of land area is agricultural land, it contributes about 17% to the total Gross Domestic Product (GDP), ten percent (10%) of total exports and offers employment to 60% of the population. India's agriculture consists of numerous crops, with the major crops of rice and wheat. Indian farmers growing pulses, sugarcane and also, non-food items like cotton, tea, coffee, and so on [3], [4].

The farmers can adapt to climate changes to some degree by shifting planting dates, choosing varieties with different growth duration, or changing crop rotations. For experimental analysis, the statistical numeric data related to agriculture is undertaken. Whereas, the clustering based techniques and supervised algorithms are utilized for managing the collected statistical data [6]. Additionally, the suitable classification methods like Support Vector Machine (SVM), neural networks are employed for better classification outcome [7]. These techniques will help in predicting the rainfall, crop yield forecasting and cost prediction of crops.

## II.CLASSIFICATION ALGORITHM FOR CROP PREDICTION

Classification which is also known as supervised learning or predictive modeling, is based on the nature of the information being extracted. Classification is a divided into two-step process – learning and classification steps. In Learning step the classification model is build using the previously known data set. In the second step which is known as classification step, if the model's accuracy is adequate then deploy this model to classify the new data.

### *Decision Tree*

Decision Tree is the one of the popular classification method that gives result in form of tree structure. Decision-tree is generally built by recursive partitioning. In this there is root and child of the tree. For the root of the tree, a single attribute split is chosen by using some criterion. For each child, the data is then divided according to the test, and the process repeats recursively. After built of the tree, a pruning step is executed, which reduces the tree size. In short, each node indicates a test on an attribute value and each branch indicates an outcome of test. It is widely used in the field of pattern recognition, machine learning and prediction. Decision tree can easily be converted to classification tree. It is very easy to understand and the provided result is worthy with small as well as large data. The data from different domain like Agriculture, Education, Medical, Diseases Analysis, Health Care, Medicine, Manufacturing, Production, Analysis of Financial, Fraud Detection and Astronomy etc. have been analyzed using Decision tree induction algorithms. The different decision tree algorithms are C4.5, ID3, CART, J48, NB Tree, REP Tree, Simple Cart and Naïve Bayes.

*J48*

The J48 algorithm was proposed by Ross Quinlan in 1993. The earlier versions are ID3 and C4.5. J48 is a classifier similar to C4.5 and C5.0. The classification tree generated is on the basis of the input attributes. The divide and conquer slant is use by decision tree. At each node, the testing of each attribute is done and the branches are prepared till leaf nodes are grasped to form a tree. The decision tree formed this way by using J48 algorithm, is an improved version of the C4.5. The pruning method is of two type. In one which is known as sub tree replacement in which few sub trees are picked up and substituted by single leaves. In the second type, a node is proceeded upwards in the direction of root of the tree by replacing other nodes through the path. It has a negligible effect on decision tree models. For the study, J48 algorithm is used as it has more accuracy rate.

*Naïve Bayes*

One of the most successful learning algorithms is Naive Bayes intended for text categorization which is based on the Bayes rule. The conditional independence between classes is the only assumption. Based on the rule, the algorithm attempts to estimate the conditional probabilities of classes given an observation. Using the joint probabilities of sample observations and classes, a simple probabilistic classifier which is based on applying Bayes' theorem having strong (naive) independence assumptions is a naive Bayes classifier. The Naive Bayes classifiers can be trained very well in a supervised learning setting which depend on the precise nature of the probability model. A benefit of the naive Bayes classifier is that it only wants a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

*Simple Cart*

The classification and regression tree algorithm was proposed by developed by Breiman et al., in 1984. The main use of CART was for data exploration and prediction. The construction of decision tree is done by using historical data set. For building up a decision tree, it is required to supply the learning sample which is a dataset of historical data having pre-allocated classes for all observations.

*SVM*

It is one of the supervised learning algorithm. It classifies the inputted data in two different classes. First of all, it makes a single or multiple hyper planes. The cases that outline the hyperplane are the support vectors. Then it chooses the hyper plane which gives the broadest way between the nearest points of the two classes. The main task of SVM is to maximize the space between the two classes to reduce the error when the given data are classified. The optimization is necessary which is only replays on the product of the pairs of sample data.

For classification, each decision tree algorithm has its particular process and all generates outcome of its classification without regardless of the outcome of rest classification algorithm. The tricky task is the selection of right algorithm as the type of data, how to retrieve the data, noisy data and time allotted to train the algorithm etc. effect on the performance and accuracy of classification algorithms[9].

## III.LITERATURE SURVEY

Hemageetha, *et al* [10] mainly focused on the soil parameters like pH, Nitrogen, and moisture for crop yield prediction. Naive Bayes algorithm was used to classify the soil and 77% of accuracy was achieved. A priori algorithm was used to associate the soil with the crops that could provide maximum yield. A comparison of accuracy achieved during classification using Naïve Bayes, J48 and JRIP is also presented.

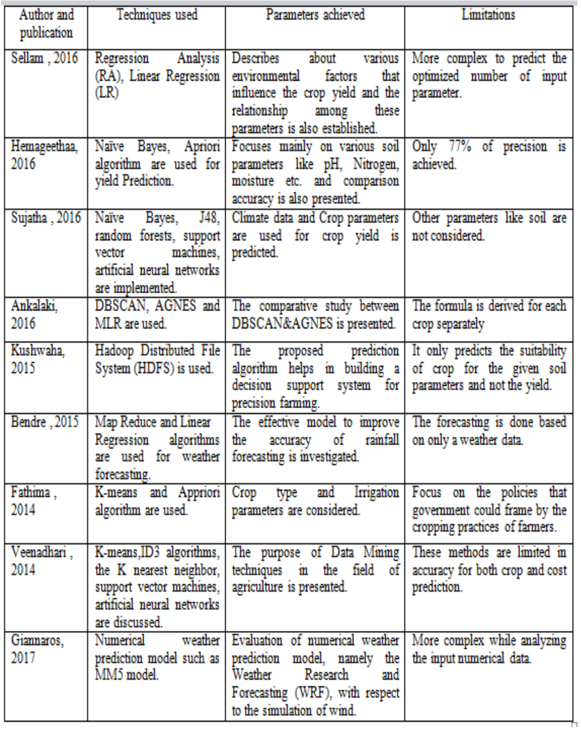
Sujatha, *et al* [11] described about the purpose of various classification techniques that could be utilized for crop yield prediction. A few of the data mining methods, such as the Naïve Bayes, J48, random forests, SVM, artificial neural networks were presented. A system using climate data and crop parameters used to predict crop growth has been proposed.

Ankalaki, *et al* [12] presented a comparative study on DBSCAN and AGNES algorithm for clustering. Crop yield was forecasted using MLR (Multiple Linear Regression) and a formula was derived for each crops. From the proposed work, we can conclude that DBSCAN was more time consuming than the optimal and efficient number of clusters. Regression analysis performed for the forecasting that showed a highly dependency on the dataset. Proper data collection will make the model significant, otherwise it can lead to inaccurate results.

Gayatri, *et al* [13] utilized IOT and web services to handle large amount of data. Sensors were used to collect the data and pass the data to data center. Agriculture field images were captured and GPS was used to accurately feed the data into repositories along with their location. Far and near nodes were communicated through cloud.

Kushwaha, *et al.* [14] predicted the suitability of a crop for a particular climatic condition and the possibilities of improving the crops quality by using weather and disease related data sets. They have proposed an analysis,classification and prediction algorithm that helps in building a decision support system for precision farming. It was based on the Hadoop file system.

*Table 1: Comparative Study of Existing System*



| Author and publication | Techniques used | Parameters achieved | Limitations |
|---|---|---|---|
| Sellam , 2016 | Regression Analysis (RA), Linear Regression (LR) | Describes about various environmental factors that influence the crop yield and the relationship among these parameters is also established. | More complex to predict the optimized number of input parameter. |
| Hemageethaa, 2016 | Naïve Bayes, Apriori algorithm are used for yield Prediction. | Focuses mainly on various soil parameters like pH, Nitrogen, moisture etc. and comparison accuracy is also presented. | Only 77% of precision is achieved. |
| Sujatha , 2016 | Naïve Bayes, J48, random forests, support vector machines, artificial neural networks are implemented. | Climate data and Crop parameters are used for crop yield is predicted. | Other parameters like soil are not considered. |
| Ankalaki, 2016 | DBSCAN, AGNES and MLR are used. | The comparative study between DBSCAN&AGNES is presented. | The formula is derived for each crop separately |
| Kushwaha, 2015 | Hadoop Distributed File System (HDFS) is used. | The proposed prediction algorithm helps in building a decision support system for precision farming. | It only predicts the suitability of crop for the given soil parameters and not the yield. |
| Bendre , 2015 | Map Reduce and Linear Regression algorithms are used for weather forecasting. | The effective model to improve the accuracy of rainfall forecasting is investigated. | The forecasting is done based on only a weather data. |
| Fathima , 2014 | K-means and Appriori algorithm are used. | Crop type and Irrigation parameters are considered. | Focus on the policies that government could frame by the cropping practices of farmers. |
| Veenadhari , 2014 | K-means,ID3 algorithms, the K nearest neighbor, support vector machines, artificial neural networks are discussed. | The purpose of Data Mining techniques in the field of agriculture is presented. | These methods are limited in accuracy for both crop and cost prediction. |
| Giannaros, 2017 | Numerical weather prediction model such as MM5 model. | Evaluation of numerical weather prediction model, namely the Weather Research and Forecasting (WRF), with respect to the simulation of wind. | More complex while analyzing the input numerical data. |

## IV CONCLUSION

There are numerous systems that utilize various methodologies to manipulate data, to derive insights and help in decision making for farmers. But the major concern is that they focus either on one crop prediction or forecast anyone parameter like either yield or price. The statistical data and predicted output are accessible for the farmers through a stand-alone user friendly application. This aids farmer to decide on the crop they would like to plant for the forthcoming year, which helps them to obtain maximum price for their products.

## REFERENCES

[1] Cervato, Cinzia, William Gallus, Pete Boysen, and Michael Larsen, "Dynamic weather forecaster: results of the testing of a collaborative, on line educational platform for weather forecasting", *Earth Science Informatics*, Vol.4, issue.4, pp.181-189, 2011.

[2] Klemm, Toni, and Renee A. McPherson, "The development of seasonal climate forecasting for agricultural producers", *Agricultural and Forest Meteorology*, Vol. 232, pp.384-399, 2017.

[3] Al-Habsi, R., Y. A. Al-Mulla, Y. Charabi, H. Al-Busaidi, and M. Al-Belushi, "Validation and Integration of Wheat Seed Emergence Prediction Model with GIS and Numerical Weather Prediction Models", In Geographical Information Systems Theory, Applications and Management, Springer International Publishing, pp. 90-103, 2016.

[4] Mughal, Muhammad Omer, Mervyn Lynch, Frank Yu, Brendan McGann, Francois Jeanneret, and John Sutton, "Wind modelling, validation and sensitivity study using Weather Research and Forecasting model in complex terrain", *Environmental Modelling & Software*, Vol.90, pp.107-125,2017.

[5] Prasad, Anup K., Lim Chai, Ramesh P. Singh, and Menas Kafatos, "Crop yield estimation model for Iowa using remote sensing and surface parameters", *International Journal of Applied Earth Observation and Geo-information*, Vol.8, issue.1, pp.26-33, 2006.

*[6]* Nayak, Ratna, P. S. Patheja, and Akhilesh A. Waoo, "An artificial neural network model for weather forecasting in Bhopal", *Advances inEngineering, Science and Management (ICAESM), 2012 International Conference on. IEEE*, 2012.

[7] Pérez-Vega, A., Travieso, C. M., Hernández-Travieso, J. G., Alonso, J. B., Dutta, M. K., & Singh, " Forecast of temperature using support vector machines", *In Computing, Communication and Automation (ICCCA), 2016 International Conference on IEEE*, pp. 388-392, 2016.

[8] Giannaros, Theodore M., Dimitrios Melas, and Ioannis Ziomas, "Performance evaluation of the Weather Research and Forecasting (WRF) model for assessing wind resource in Greece", Renewable Energy, Vol.102, pp. 190-198, 2017.

[9] Sellam,V, Poovammal, E., "Prediction of Crop Yield using Regression Analysis", Indian Journal of Science and Technology, Vol. 9, issue.38,pp.1- 5, 2016.