Crop Yield Prediction & Crop Selection using Machine Learning

Janki A. Patel[1], Sachi R. Patel[2] and Dr. Maulika S. Patel[3]
[1]Computer Engineering, G. H. Patel College of Engineering, India
[2]Computer Engineering, G. H. Patel College of Engineering, India
[3]Head of Department, Computer Engineering, G. H. Patel College of Engineering, India

Abstract

Agriculture sector being prevalent employer in India's economy, the declining rate of its share GDP (17.32%) is of imperative concern. Moreover, with limited amount of resources like land, it becomes the responsibility of technology to provide methods which can make judicious use of the resources. Thus, we are proposing a system that serves above issue and which can be helpful to the people new in the field of agriculture and do not have basic knowledge of what should be grown when. The system provides two modules: Crop Selection and Crop Yield Prediction. It will not only address to problems of both neophyte and veteran farmers, but also to the government. Depending upon the production of crop, government can decide various policies, minimum support price of crop and keep track of the flow of supply in the market. Firstly, while selection of crop is very crucial decision in each season for a farmer, it must be done scrupulously. After the user selects the season of cultivation, list of cultivable crops will be diminished to the list of crops favorable to that season. A selecting factor will be calculated of the enlisted crops based on the product of minimum support price & predicted yield and two topmost crops having maximum selecting factor will be given as output. Furthermore, dataset is formed by aggregating different datasets available on Indian government website for data. Secondly, the yield (production in tones/area in km^2) prediction was obtained by implementing regression algorithms like linear regression (with multiple variables) and SVM on dataset with attributes like District, Area, Production, Season, Crop, Rainfall and Year. Regression algorithms divide the datasets into two parts- Training dataset and testing dataset. The RMSE (Root Mean Square Error) for each algorithm was calculated in which, SVM Liner kernel turns to be most appropriate for the purpose with RMSE of 0.42. Summarizing, after taking area of cultivation as input, the crop yield will be given as output of SVM linear kernel algorithm. The platforms used for the system are: i) R for building model of regression algorithms and ii) WampServer and PHP for online portal.

Keywords: *Crop Suggestion, Crop Yield, Machine learning, Precision Agriculture, Prediction, Regression Algorithm*

Author for Correspondence E-mail: maulikapatel@gcet.ac.in, Tel: +91 9428488563

INTRODUCTION

Agriculture is one of the main building blocks of Indian economy. Numerous disciplines have been evolved with an intention contributing to the field of agriculture. Though majority of framers faces many issues including how, where, and when to irrigate, fertilize, etc. Because of such issues, crop production often gets affected & farmers do not get the required or expected outcome. Thus in an effort to aid farmers, a new branch known as Precision Agriculture or Precision Farming has been evolved. It is one of the many modern farming practices that generate accurate results and makes the production more efficient. There are various cases when due to performing wrong task at wrong time results into a major loss in form of less productivity or nutrition deficient grain. This is the case for farmers whose main resource for money is only agriculture and are well experienced. If an experienced farmer is prone to make wrong decision, how can one expect that someone new to this field is entirely aware of the imminent hurdles in their journey?

Supporting our purpose, the objective is served accordingly. By monitoring soil, crop and climate in a field and providing decision support system that is able to learn, it is possible to deliver treatments, such as irrigation fertilizer and pesticide application. Developing better techniques to predict crop productivity in different climatic conditions can assist farmer in better decision making. Thus, our aim here would be to guide farmers in taking some immensely important decision by studying different parameters involved in it so that they will not suffer from huge losses which affect their livelihood. The research focuses on providing support to the farmers in the form of a tool which will help them in deciding various parameters involved in farming process. Decisions such as when, where & how much fertilizer, irrigation, etc. must be applied will be determined. *It* can help farmers know how much and when to apply these inputs. Moreover, crop yield prediction and best suited crop suggestion will be provided. However, *Precision agriculture* relies upon specialized equipment, software and IT services [14].

More specifically, farmers can implement according to directions provided in their fields. Also, the government can use it to guide farmers who cannot afford computer or are still deprived from facilities such as electricity. For assessing productivity of crops, one should know about various indices related to agriculture. Example: NPP (Net Primary Productivity), NDVI (Normalized Difference Vegetation Index), VCI (Vegetation Condition Index), TCI (Temperature Condition), etc. Using rainfall data and surface temperature data we can predict the yield of crops respective to crop & climatic conditions [4]. It requires different datasets to study and identify patterns to find above mention indices and take various decisions.

LITERATURE REVIEW

N. Gandhi *et al.* [1] has in her paper, examined various data mining techniques in order to predict yield for rice crop in Maharashtra state, India. They acquired dataset of 27 districts with attributes like Precipitation, area, production, evapotranspiration, minimum temperature, maximum temperature, normal temperature, range, generation and yield for Kharif season (from 1998-2002) from publicly available Indian Government datasets. They used WEKA tool as data processing. However, the prediction was done in 3 classes: high, moderate and low with an accuracy of 82.51%.They extended the work done in [1] in [2] as classifiers used before were SMO and SVM. While using classifiers BayesNet and NaiveBayes, the outcome showed that BayesNet was much better than NaïveBayes. Accuracy achieved was: BayesNet-97.53% & Naivebayes-84.69%.

Karandeep Kaur *et al.* [3] proposed various applications of Machine Learning that can be useful in the field of agriculture such as Crop selection & Crop yield prediction, Weather Prediction, Smart irrigation system, Crop disease prediction and in decision making of minimum support price of crops. Also, she presented appropriate algorithms required to accomplish the applications represented by respective authors. However, no method was mentioned in the paper. Nishit Jain *et al.* [4] has worked on implementing a better crop selection system which takes in consideration different factors which play important role in selection and yield of the crop. Factors like soil type and its chemical composition, temperature, minimum price and total production are deciding parameters in the two modules: Crop selection & Crop Sequencing. However, predicted yield required in selection method is a result of classification and to perform ranking precisely, it is advisable to predict numeric value of yield i.e. regression algorithms.

Aakunuri Manjula *et al.* [5] made use of vegetation indices collected through remote sensing technology, climate associated variables, weather disturbance data and agronomic variables to propose framework for crop yield prediction. The vegetation indices include Normalize Difference Vegetation Index (NDVI), Vegetation Condition Index (VCI) and Temperature Condition Index (TCI). It has provision for crop selection, independent variable selection, dependent variable section, dataset selection, pre-processing and crop yield prediction. It allows flexibility to have different

combinations of inputs and outputs. Similarly, Ma Qingyuan et al. used remote sensing data to retrieve LPP, NDVI, LAI, etc. Also, calculated crop evapotranspiration and Soil adjusted vegetation index (SAVI) and performed supervised classification on spatial data.

DATASETS

The datasets used for training and testing purpose for predicting yield of the crops were generated using following publicly available dataset:

Rainfall dataset
A dataset of rainfall in cm, for each district of each state for 10 years. Attributes -State, District, Kharif, Rabi, Summer, Winter and Whole Year.

Production dataset
It contains dataset of the production of particular crop in determined area, district-wise. Only primary crops which are having favourable climatic conditions with respect to area are available in the data. Attributes - State, District, Crop, Season, Area, Production and Year.

The crop and season are selected for which prediction is supposed to be calculated. According to crop and season, attributes from rainfall dataset and production dataset are selected. Using R tool, dataset gets generated by running an SQL group-by query which matches the State, district, year and season from both of the datasets. Finally, the dataset will have attributes: State, District, Year, Crop, Season, Area, Production, and Rainfall.

METHODOLOGY

Yield monitors can predict the amount of yield variability at a smaller scale. This can form guidance to each separate area rather than whole area. The principle procedure is firstly divide the dataset into training and testing dataset in a predefined ratio such as 75:25, that is from the entire dataset, 75 percent of data is used for training and the rest 25 percent of data will be used for testing purpose. Next, the regression model which implement the various regression algorithms to perform actual computations is built.

Crop Yield Prediction
Crop yield is the production of a particular crop per area. The yield which is to be predicted is a numeric value, in order to do so, machine learning concepts have been used.
Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed [6]. There are two machine learning algorithms used for prediction, well known as supervised learning and unsupervised learning. Both consist of wide area of algorithms which are applicable for either predicting a known class label or predicting an unknown label respectively. Specifically in this research area regression algorithms of supervised learning are implemented. Following regression algorithms are implemented using R:

***Linear Regression*:** Simple linear regression is a type of regression which has two variables, criterion variable and predictor variable. Criterion variable is the variable which we are predicting whereas predictor variable is the one on which we are basing our prediction. When there are more than one predictor variables, it is known as linear regression with multiple variables. It is a statistical method that allows us to summarize and study relationships between continuous (quantitative) variables [7].

In our study we have four predictor variables (also known as features):

- x_1 - area
- x_2 - production
- x_3 – season

- x_4 – rainfall

and criterion variable y – yield

For single predictor, hypothesis takes the form:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

We are having multiple features so hypothesis takes the form:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4$$

The parameter theta is determined by the cost function [8]. Linear regression tries to find best fitting straight line through the points, from this line the error on prediction is always least[7].

***Support Vector Regression*:** Support Vector machine is a supervised algorithm which can be used for both classification as well as regression.

In Support Vector Regression (SVR), given a training data set $D = \{(x_1, y_1), \ldots, (x_M, y_M)\}$ of M samples, where $x^i \in R^N$ are multi-dimensional inputs and $y_i \in R$ are continuous uni-dimensional outputs, we seek to find a continuous mapping function $f : R^N \rightarrow R$ that best predicts the set of training points with the function $y = f(x)$[10].

SVR is utilized to resolve problems which are not linearly separable, that is problem which cannot be separated by straight line. In order to solve this SVM offers various kernel functions that transform nonlinear spaces into linear one. The three kernels used in present study are linear, polynomial and radial. The kernel functions transform the data into a higher dimensional feature space to make it possible to perform the linear separation [11]. All these are implemented using CRAN and e1071 packages available in RStudio.

Crop Selection

For crop selection one has to provide the season for which crop is to be selected and on the basis of the selected season as well as predicted yield, the most profit-making crop will be suggested. A website interface is designed at the front end, where at the back end following process takes place:

Before selecting crop, these two steps are performed at back end:

i). Crop classification

ii). Crop Sequencing

Crop classification: Here, crops are classified according to the season, i.e. maximum yield producing season. The groups formed there by, are further used by Crop sequencing method for suggesting the best crop among all, for that particular season.

***Crop sequencing*:** Each classified branch is sequenced, i.e. the crop having maximized select factor at first position.

Select Factor = Net Yield * Price

Net Yield: It is predicted for each crop using attributes such as Precipitation, area, season and yield

***Price*:** It is market price of the respective crop. The price factor is one of the most important factors which play a major role in selecting crop. For example, there are two crops and both produce equal yield but one crop is valued at a lower price than the other. If the price factor is not included in the crop selection method, then system may lead to select a wrong crop to grow. Thus, according to the input (season), the foremost crop in the sequence (most productive crop) is given as the output.

RESULTS

Regression algorithm being used produce results in a numeric form, gives quantitative yield prediction of crops, by making use of various built-in packages accessible in R for classification and regression. Once the yields are enumerated, the accuracy of an algorithm is considered on the basis of Root Mean Squared Error (RMSE).

RMSE of an algorithm indicates how close the observed data points are to the model's predicted values [12]. It is the square root of the variance of the residuals.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

Here p_i is predicted value and a_i is the actual value [13]. This value indicates correctness of the outcome and thus makes it viable as well as practical for choosing the leading model for particular task.

***Table 1**: RMSE of Regression Algorithms*

Regression Algorithm	RMSE
Logistic Regression	0.8277
SVR: Linear kernel	0.4282
SVR: Polynomial kernel	2.1676
SVR: Radial kernel	0.4680

The model with least error is considered best and for current research Support Vector Regression with linear kernel gives minimum error. Thereby, linear kernel regression algorithm gives better results compared with all other regression model, which can be concluded from Table 1.

CONCLUSIONS

These same methods can be applied in order to determine how much & when to irrigate, given that we have the dataset with attributes Production, area, yield, month, when and water consumed. On the other hand one can also govern how much & when to fertilize by applying various regression algorithms.

We thereby, conclude that the proposed work is feasible and can be applied for real world applications. The machine learning algorithms can be beneficial in order to obtain efficient results by analyzing various pattern from historical data. Government as well as farmers can thus use these system for taking decisions.

ACKNOWLEDGMENTS

Acknowledgments recognize the contribution of funding bodies and anyone who has assisted in the work.

REFERENCES

1. Petkar O, Tripathy A. Rice Crop Yield Prediction in India using Support Vector Machines. In: Gandhi N, Armstrong L. (Eds.). *The 13thInternational Joint Conference on Computer Science and Software Engineering (JCSSE)*; 2016 July 13-15; Khon Kaen, Thailand. India: IEEE; 2016. 5p.
2. Petkar O. Predicting Rice Crop Yield using Bayesian Networks. In: Gandhi N, Armstrong L. (Eds.). *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*; 2016 Sept 21-24; Jaipur, India. India: IEEE; 2016. 795-5p.
3. Kaur K. Machine Learning: Applications in Indian Agriculture. *International Journal of Advanced Research in Computer and Communication Engineering.* 2016; 5(4): 342-3p.

4. Jain N, Kumar A, Garud S, Pradhan V, et al. Crop Selection Method Based on Various Environmental Factors Using Machine Learning. *International Journal of Advanced Research in Computer and Communication Engineering*. 2017; 4(2): 1530-4p.
5. Narshimha G. XCYPF: A Flexible and Extensible Framework for Agricultural Crop Yield Prediction. In: Manjula A (Ed). *IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO)*; 2015 Jan 9-10; Coimbatore, India. India: IEEE; 2015. 5p.
6. Machine Learning. *Wikipedia, the free encyclopedia* [Internet]. 2003 May 25 [cited 2018 April 29]; Available from: https://en.wikipedia.org/ wiki/Machine_learning
7. Introduction to Linear Regression. *Online Statistics Education: An Interactive Multimedia Course of Study* [Internet]. 2004 Nov 14 [cited 2018 April 30]; Available from: http://onlinestatbook .com/2/regression/intro.html
8. Linear Regression with Multiple variables. *Notes of the Stanford's machine learning course* [Internet]. 2015 Nov 25 [cited 2018 May 18]; Available from: http://www.dmi.unict.it/farinella/SMM/Lectures/25_Nov2015_2.pdf
9. Goumopoulos C. Applying Machine Learning to Extract New Knowledge in Precision Agriculture Applications. In: Dimitriadis S(Ed). *Panhellenic Conference on Informatics*; 2008 Aug 28-30; Samos, Greece. India: IEEE; 2008.100-5p.
10. Advanced Machine Learning Practical 4: Regression (SVR, RVR, GPR). *Learning Algorithms and Systems Laboratory* [Internet]. 2016 May 5 [cited 2018 May 12]; Available from:http://lasa.epfl.ch/teaching/lectures/ML_MSc_Advanced/Practical/tp4_regression_sol.pdf
11. Support Vector Machine – Regression (SVR). *An Introduction to Data Science* [Internet].2007 June 28 [cited 2018 May 12]; Available from: http://www.saedsayad.com/support_vector_machine_reg.htm
12. Assessing the fit of Regression models. *THE ANALYSIS FACTOR* [Internet]. 2012 Jul 3[cited 2018 May 13]; Available from: https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/
13. Model Evaluation – Regression. *An Introduction to Data Science* [Internet]. 2007 June 28 [cited 2018 May 13]; Available from: http://www.saedsayad.com/model_evaluation_r.htm
14. Precision Agriculture.*WhatIs.com [*Internet]. 2016 Feb 15 [cited 2018 May 18]; Available from: https://whatis.techtarget.com/definition/precision-agriculture-precision-farming
15. Preethaa K, Nishanthini S, Santhiya D, et al. Crop yield prediction. *International Journal on Engineering Technology and Sciences-IJETS*. 2016; 3(3): 111-6p.
16. Misra B, Singh C. Machine learning approach for forecasting crop yield based on climatic parameters. In: Veenadhari S (Ed). *International Conference on Computer Communication and Informatics (ICCCI)*; 2014 Jan 3-5 2014; Coimbatore, India. India: IEEE; 2014. 5p.