Advances and Applications in Mathematical Sciences
Volume 20, Issue 11, September 2021, Pages 2803-2810
© 2021 Mili Publications

A MACHINE LEARNING APPROACH FOR CROP YIELD PREDICTION IN AGRICULTURE

S. RAO CHINTALAPUDI, CHANDRA SEKHAR KOPPIREDDY and VIJAYA KUMAR GUBBALA

Department of CSE
Pragati Engineering College
(Autonomous) Surampalem
Andhra Pradesh, India
E-mail: srao.chintalapudi@gmail.com
sekhar222.k@gmail.com
vijay9908914010@gmail.com

Abstract

In agriculture sector, Crop Yield Prediction is an important issue to address food security challenges and reducing the impact of climate change. In this paper, machine learning techniques are used to predict yield of most consumed crops using publicly available data from FAO and World Data Bank. Different Regression analysis algorithms such as Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor are applied on the dataset to predict the crop yield. The performance of these algorithms is measured using R^2 score and among these three algorithms Decision Tree Regressor results good R^2 Score.

I. Introduction

Agriculture plays a critical role in the global economy. With the expansion of human population, understanding worldwide crop yield is an important issue to address food security challenges and reducing the impact of climate change. The agricultural Yield is primarily depends on usage of pesticides, weather conditions like rainfall, temperature [1] [2] etc. Accurate information about history of crop yield will play a crucial role for future yield prediction. Climatic factors include humidity, rainfall, temperature and

2010 Mathematics Subject Classification: 68T45, 62Jxx.
Keywords: Crop yield prediction; machine learning; regression analysis; decision tree regressor.
Received October 12, 2020; Accepted November 4, 2020

Environmental factors like pesticides usage, soil condition will greatly affect the yield of crop. In this paper, author considered two climatic factors namely rainfall, temperature and one environmental factor namely usage of pesticides for building a better model.

Machine Learning can be used in sectors ranging from health care to Agriculture. In Agriculture sector, machine learning is being used for several years. The most studied problem in agriculture is Crop Yield Prediction and several models were developed to predict yield of a particular crop so far. Even though, the latest crop prediction models can predict yield effectively. There is need of robust model that predict yield of a Crop by considering all the factors that impact crop yield [6].

The remainder of this paper is organized is as follows: section 2 explains the related work. Section 3 discusses Regression Analysis for Crop Yield Prediction. The experimental setup and about the dataset is explained in section 4. Results and Discussion is presented in Section 5. Finally, section 6 concludes the paper.

II. Related Work

Japneet Kaur [3] in his paper studied four Indian crops such as Rice, Cotton, Wheat and Sugarcane and this state level data collected for the period 2004 to 2013. Different climatic conditions of seven agricultural based states data is studied and analysed to predict the impact of climate change on crop yield. Rasul G et al. [1] studied impact of sudden increase in temperature on yield of a crop. In his paper, he also studied the effect of sudden rise in temperature on country wide agriculture issues. Pratap S. Birthal et al. [4] also studied agriculture based countries like India, where agriculture land is limited but people are highly dependent on agriculture. In his paper, he also discussed about impact of natural calamities such as floods, cyclones on crop yield. These natural calamities will decrease the agricultural productivity, that in turn lead to food security issues. It is also suggested that countries need to improve their technological and financial capabilities to overcome the impact of climate change. Yunis H et al. [5] investigated that how bacteria grows in a particular temperature range and how the bacteria spoil the yield. He experimented on tomato crop and studied the impact of bacteria on crop yield. He also studied the correlation between temperature and growth of bacteria. His research revealed that plant disease will spread in the temperature between 13 and 280C.

III. Crop Yield Prediction

Regression Anlaysis is a form of predictive modeling technique to find out the relationship between dependent and independent variables of the data. Three regression models [6]-[12] are used to predict crop yield namely Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor.

A. Decision Tree Regressor (DTR). Decision Tree algorithm is one of the most widely used machine learning algorithm, due to the fact that it will work well with noisy or missing data and can easily ensembled to form more robust predictors. Decision Tree can be used both in classification and regression problem.

Decision Tree Regression use Mean Squared Error (MSE) to decide to split a node in two or more sub-nodes. The attribute at root node will split the data into subsets and this process will continue up to leaf node or decision node reached. Attributes in the nodes are at different level in different decision trees. Weighted average (i.e. MSE*num_samples) of two new nodes is considered to find out the best split. Best split can be done based on trail and error method that means each and every attribute will tried and compute the score. This process will be repeated for all the nodes until stopping condition is met such as max_depth.

B. Random Forest Regressor (RFR). Random Forest is an ensemble method that utilizes Decision tree algorithm. This method can be used for both regression task and classification task. A Random Forest uses multiple decision trees and performs averaging the prediction of trees. The following two points are core of the random forest algorithm.

- Training samples are picked randomly.
- Random subsets of features for splitting nodes.

C. Gradient Boosting Regressor (GBR). Gradient Boosting Regressors (GBR) is an ensemble method that uses decision tree algorithm.

For each decision tree, it computes the difference between predicted value and actual value, called residual. Model will be trained in such a way that the residual should be minimum. In this way, model performance can be improved and produce good results.

The parameter n_estimators is taken as 200 for this problem indicates number of trees in the forest. If n_estimators is more than the model can learn data in a better way. In contrary, if number of estimators increases, then the runtime of the algorithm is also increases. Here, the value of the parameter max_depth is 3 which is a default value for Gradient Boosting Regressor.

IV. Experimental Setup and Dataset

A. Experimental Setup. Crop Yield Prediction models were implemented in Google Colab which is a cloud based jupyter notebook. Python 3 is used as a programming language. To use regression algorithms like Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regresser, the popular library for machine learning- sklearn is used. R^2 score is computed using the function R^2 score in the sklean. metrics library.

B. Dataset. The Dataset for crop Yield Prediction is collected from Food and Agriculture Organization of the United Nations (FAOSTAT) website [13]. It offers national and international statistics on food and agriculture. The Yield dataset collected from FAOSTAT contains the attributes like Domain code, Domain, Area Code, Area, Element code, Element, Item Code, Item, Year code, Year, Unit, Value. The yield data is available from the year 1961 to 2016.

Rainfall has a dramatic effect on agriculture, so rainfall per year information was downloaded from the World Data Bank [14] in addition to average temperature for each country. The final data frame for average rainfall includes country, year and average rainfall per year. The temperature data is available from 1743 to 2013. Data for pesticides was collected from FAOSTAT. It is noted that the pesticides data available from 1990 to 2016.

C. Dataset Gathering and Preprocessing. Crop Yield data is downloaded from FAOSTAT and the columns that won't be of any use analysis such as Domain code, Domain, Area Code, Item Code, Year code are dropped from the dataset. Also the column Value is renamed as hg/ha_yield to make it easier to recognize that it is crop yield production value in Hectogram per Hectare. The final dataframe consists of only four fields such as country, item, year, hg/ha_yield. Merging all the three dataframes together, the year range will start from 1990 and ends in 2013, that is 23 years worth of data.

If the data is gathered from different sources and collected in raw format, then it is not feasible for analysis. Hence, one has to convert the raw data into clean dataset using data preprocessing techniques. The final dataset has two categorical columns such as Items, Countries. Because of many machine learning algorithms cannot operate on label data directly, those two columns are converted into numerical form using one hot encoding. Now, the dataset has 115 features which are highly varying in magnitudes. By applying minmax scalar, all the features of the dataset are in same scale that is [0, 1]. Finally, the dataset is split into training and testing data and the training data is a given as input to the machine learning algorithm. Now, exploring the relationships between the columns of the dataset, a good way to quickly check correlations among columns is by visualizing the correlation matrix as a heat map. It is evident from the heat map in Figure 1 that all of the variables are independent from each, with no correlations.

V. Results and Discussion

In order to measure the performance of these three regression models, the metric R^2 score is used. The R^2 score [15] is defined as the proportion of the variance in the dependent variable that is predictable from the independent variable. It is a statistical measure between 0 and 1 which calculates how similar a regression line is to the data it is fitted to. Another definition for R^2 score is given using Equation (1).

$$R^2 Score = \frac{Total\_Var\;iance\_\exp\;lained\_by\_\mod\;el}{Total\_Var\;iance} \tag{1}$$

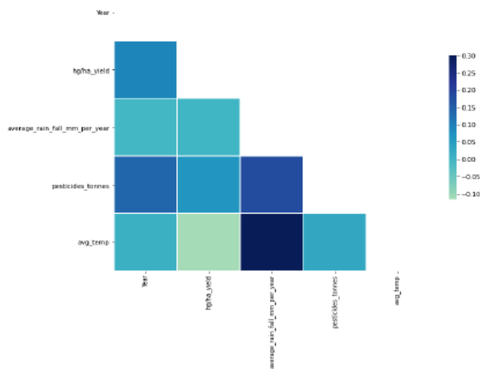

Figure 1. Heat map that represents Correlation between features in the dataset.

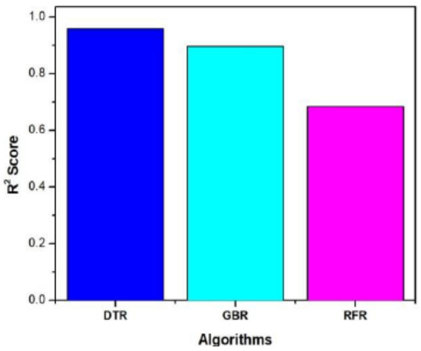

Figure 2. Comparison of three models (DTR, GBR, RFR) in terms of R^2 score.

If it is 100% the two variables are perfectly correlated that is with no variance at all. A low value would show a low level of correlation that indicates poor regression model. The most common interpretation of R^2 score is how well the regression model fits the observed data. The final dataset is divided into Training set and Testing set in 70:30 proportion and given as input to the three algorithms such as Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor. The performance of these regression models is computed in terms of R^2 score. The R2 score of Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR) is 0.96, 0.90 and 0.68 respectively and the results are depicted in Figure 2. From the results shown in Figure 2, Decision Tree

Regressor has the highest R^2 score of 96%, Gradient Boosting Regressor comes second with 89%. For example, an R^2 score of 70% says that 70% of the data fit the regression model. Generally, a higher R^2 indicates a better fit for the model. From the obtained results, it’s clear that the model fits the data to a very good measure of 96%.

VI. Conclusion and Future Scope

In this paper, Machine learning Techniques are applied to solve one of the essential task in agriculture sector i.e. Crop Yield Prediction. Crop yield prediction models will help the farmers to take an appropriate decision regarding selection of crop to grow. Crop yield information and Pesticides information is collected from FAOSTAT, rainfall information and average temperature data is collected from World Data Bank. After merging the data from the different sources, one hot encoding is applied on Country and Item attributes to convert them into numerical form. To avoid varying magnitudes of the values scaling has been done on the dataset using Min Max Scalar. The final dataset is divided into Training set and Testing set in 70:30 proportion and given as input to the three algorithms such as Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor. The performance of these regression models is computed in terms of R^2 score. The R^2 score of Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR) is 0.96, 0.90 and 0.68 respectively. From the obtained results, it is clear that the Decision Tree Regression model fits the data to a very good measure of 96%. In future, the accuracy of the prediction can be further improved using deep neural networks and by adding additional features like soil information, solar information availability of other water resources and so on.

References

[1] G. Rasul, Q. Z. Chaudhry, A. Mahmood and K. W. Hyder, Effect of. 28- 40Temperature Rise on Crop Growth and Productivity, Journal of Meteorology 8(15) (2011), 7-8.

[2] Anupama Mahato, Climate Change and its Impact on Agriculture, International Journal of Scientific and Research Publications, ISSN 2250-3153 Volume 4(4) (2014), 4-5.

[3] Japneet Kaur, Impact of Climate Change on Agricultural Productivity and Food Security

Resulting in Poverty in India, Università Ca' Foscari Venezia 23 (2017), 16-18.

[4] S. Pratap Birthal, Md. Tajuddin Khan, Digvijay S. Negi and Shaily Agarwal, Impact of Climate Change on Yields of Major Food Crops in India: Implications for Food Security, Agricultural Economics Research Review 27(2) (2014), 145-155.

[5] H. Yunis, Y. Bashan, Y. Okon and Y. Henis, Weather Dependence, Yield Losses, and Control of Bacterial Speck of Tomato Caused by Pseudomonas tomato, American Phytopathological Society (1980), 1-2.

[6] J. P. Powell and S. Reinhard, Measuring the effects of extreme weather events on yields, Weather and Climate Extremes 12 (2016), 69-79.

[7] G. P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50 (2003), 159-175.

[8] B. Dumont, V. Leemans, Salvador Ferrandis, Bernard bodson and Jean-Perrie Destain, Assessing the potential of an algorithm based on mean climatic data to predict wheat yield., Precision Agriculture 15(3) (2014), 255-272.

[9] B. Basso, B. Bodson, V. Leemans, B. Bodson and J. P. Destain, M-F Destain, A comparison of within season yield predictions algorithm based on crop model behaviour analysis, Agricultural and Forest Meteorology 204 (2015), 10-21.

[10] A. Stanley Changnon, Prediction of corn and soya bean yields using weather data, CHIAA Research Report No. 22, Crop-Hail Insurance Actuarial Association (1965), 6-10.

[11] J. Betty, G. Shem Juma and O. Everline, On the Use of Regression Models to Predict Tea Crop Yield Responses to Climate Change: A Case of Nandi East, Sub-County of Nandi County, Kenya, Assesing the Value of Systematic Cycling in a Polluted Urban Environment, Climate 5(3) (2017), 5.

[12] Christian Baron, Mathieu Vrac, P. Oettli, B. Sultan, Are regional climate models relevant for crop yield prediction in West Africa, Environmental Research Letters 6 (2011), 2-6.

[13] FAOSTAT website [online] Available: http://www.fao.org /home/en/

[14] World Data Bank Website [online] Available: https://data.worldbank.org/

[15] Coefficient of Determination, In: The Concise Encyclopedia of Statistics. Springer, New York, NY. 2008 https://doi.org/10.1007/978-0-387-32833-1_62