Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776946

A MACHINE LEARNING TECHNIQUE TO PREDICT AGRICULTURAL YIELD

[1]Deevi Radha Rani, [2]Ch V Phani Krishna, [3]Venkata Rami Reddy Ch, [4]K. Swetha, [5]K Narasimha Raju, [1]Sajja Radharani

[1]Department of CSE, VFSTR Deemed to be University, Guntur, India
[2] Department of CSE, TKR Engineering College, Hyderabad, India
[3]School of Computer Science and Engineering, VIT-AP University, India
[4]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
[5] Department of CSE, Lendi Institute of Engineering and Technology, Vizianagaram, India

Abstract

India is an agriculturally dependent nation, and as such, agriculture both totally and partially determines the nation's economic situation. Seasonal and economic factors have an impact on agriculture. It is difficult to anticipate crop yields since one must use the data that are currently available. This study focuses on applying machine learning techniques to predict crop yield because they are a critical decision-supporting tool. Before planting their crop, farmers will find this document useful in determining crop output. Various supervised machine learning approaches that can be used for prediction. In this study, regression methods including Decision Tree Regression, Multiple Linear Regression, and Random Forest Regression are used.

Keywords - Decision Tree (DT), Machine Learning (ML), Multiple Linear Regression (MLR), Random Forest Regression (RFR)

I. Introduction

Agriculture is backbone of our economy [11]. In our country due to increasing population of food, progress in the agriculture sector is required to meet the demands. From past times agriculture is important and important culture practiced in India [14]. Ancient people grow crops on their lands and have been adapted to their needs and here it is normal plants can grow and used by many creatures such as animals, humans and birds. Since the invention of new technologies and techniques, the field of agriculture is slowly falling. Due to this fact many inventors have focus on growing hybrid synthetic products that leads to unhealthy lives. Today modern people don't know how to grow plants at the right time and in the right place it's because of climatic conditions also change towards basic resources like soil, water, air and lead to food.

There is no solution and technology to overcome the situations we are facing. There are many opportunities for increasing economic growth in the agricultural sector. There are different ways to go and improve crop yield and quality. Data Mining is useful for forecasting crop yield production [14]. Data mining software is an analysis tool it enables users to analyze data from different dimensions [8]. Data Mining is one of the processes of extracting and discovering patterns in the large relational datasets [4]. By combining all the data it provides the information. These information is transformed into historical patterns and feature trends. For suppose total information about plant production can help farmers identify and prevent future crop losses. Forecasting crop yield is an important agronomic matter. Every Farmer always tries to find out how much return they will yield from his forecast. The main goals are:

- Using machine learning techniques to predict crop yields.
- To provide an easy to use user interface.
- Increased accuracy of crop yield forecasts.
- Analysis of various climate parameters.

Most of agricultural crops have several degrade in terms of performance over the past two decades. This paper will help farmers to know the yields of their crops before planting in the agricultural field. It helps them to keep right decisions [4] [6]. An attempt was made to solve the problem by building a prototype of the interactive prediction system. Such a system is implemented with an easy-to-use machine learning algorithm [3] [15] [18]. Forecast results are provided to the farmers therefore different techniques for this type of data analysis exist in forecasting the harvest and with the help of these algorithms we can predict harvest yield.

II. Literature Survey

The Literature has many reported walks in this domain. Dahikar, S.S, et al. proposed a model by taking the account a variety of the climatologically related circumstances that have an impact on local weather conditions in the different parts of the world [1]. Based on weather conditions crop yield is influenced. The parameter of their regional soil parameter is included. The feed forward back propagation ANN is then used to analyse it. To make the ANN approach more effective, analyse it in Mat lab.

In [2] authors employ advanced regression methods such as ENet algorithms, Kernel Ridge, Lasso the principle of Stacking Regression to improve the algorithms. When those models were used independently, the result was much more improvised. The output depicted in the figure is currently a web application; however, our future work will focus on developing an app that farmers can use and translating the entire system into their native language.

In [3] machine learning techniques for finding the accurate crop yield prediction and estimation of nitrogen status have been developed over the last 15 years. According to the paper, rapid advanced in sensing technologies. The machine learning methods provides cost-effective and detailed solutions. It helps in decision making and improving crop and environmental state estimation.

In [4], authors have gone over a number of algorithms related to data mining classification methods. These algorithms are applied to dataset that has been compiled over time in order to predict soya bean crop yields. As a result, the bagging algorithm of the classification technique is well suited for predicting the soya bean crop yields.

Bondree et al. implements a system that uses past data to predicting the crop yield. This is done by using Machine Learning Algorithms those are Support Vector Machine and Random Forest is to analyse agricultural data and it suggest the best fertilizer for each crop. The purpose of this paper is to develop a prediction of the crop yield model it can be used in the future generations [5].

In [6] authors had constructed a model to test the soil fertility. Based on the sensor, they obtained a value which is used to suggest the best crop for planting. In predictive analysis they followed the two models they are predictive model and descriptive model. This application is mainly used to test the soil fertility and suggest the best suitable crop. This paper is used to suggest the best fertilizer for the soil and it will increase the profit of that yield.

The population in India depends on the agriculture for livelihood, the agriculture output that depends on the climatic conditions, weather and soil conditions. The parameters used in the paper are temperature, rainfall, location and soil conditions to predict the crop suitability. The different types of techniques used are decision tree, K-Nearest Neighbours, Neural networks and Map visualization. By using this system, the farmers can be able to make an informed decision on growing the particular type of crop variety based on the geographical and environmental conditions [7].

Vijay Baskar, et al. [8] had constructed a model to test the soil fertility. Based on the sensor, they obtained a value which is used to suggest the best crop for planting. In predictive analysis they followed the two models they are predictive model and descriptive model. This application is mainly used to test the soil fertility and suggest the best suitable crop. This paper is used to suggest the best fertilizer for the soil and it will increase the profit of that yield.

There are various types of parameters for crop selection to maximize crop yield rate. The parameters are market price, production rate and policies. Based on the studies of many researchers we can able to predict the crop yielding rate, classification of soil and crop, weather predictions for agriculture planning using the statistical models. They are different types of techniques like the SVM (support vector machine), Gradient boosted decision tree, K-Nearest are used. In this paper the CSM technique is presented for selecting the crop sequence which will be planted over the seasons. By using this proposed method, a solution for crop selection based on yield rate prediction is obtained by involving the parameters like water, weather, type of soil etc... The predicted value whichis obtained from the parameters is dependent ontheperformance and accuracy of CSM method [9].

In [10] authors had discussed that there is the major issue where most of farmers label by precision agriculture. Indian farmers they don't choose right crop according to their soil fertility. Because they will face a. Farmers can plant the yield which increase their profit. In this paper, they used some of the proposed techniques like decision tree, k-nearest and also naive base for improving accuracy and its efficiency.

III. Current Solutions

Simple Linear Regression Technique is discussed in[8][10]. “Multiple Linear Regression [14] is used to find the relationship between a single input variable and output variable [12] [15]. This relationship shows how an input variable was related output variable [16] [20].”

It has two main objectives.

1. To establish the relationship between the two variables if there is exist. More specifically we will establish the statistical significant relationship between the two variables.
2. To Forecast new observations

“Multiple Linear Regression is used to estimate the relationship between two (or) more independent variables and a single Dependent variable”. The generated models were dependent on more than one feature then there is increase in getting better-fit. Multiple Linear Regression can detect the outliers more efficiently. When the Linearity of dataset decreases then the accuracy will decreases. As we use all features to generate the model, then there is a over fitting problem.

“Random Forest Regression is another type of supervised machine learning technique which is very effective[2][3]. Before we jump on to the [5] random forest Regression let’s first understand the concept of ensemble learning [11]. In this technique which takes predictions

from multiple machine learning algorithms or predicting for same algorithms multiple times in order to get more accurate results".

Predicting the single individual model that will not give accurate results and hence we can use an ensemble learning techniques [7]. So we can develop several models like KNN and support vector machines [11] [19]on same data to get the combined predictions from them similarly. We can combine several decision trees that are trained together to get final prediction.

Ensemble learning comes in two ways, they are: Boosting, Backing. Boosting learners are learnt sequentially with early learners by fitting simple model to the data and then analysing data for errors. Consecutive trees or random samples are fed and at every step the goal is to improve the accuracy from the prior tree.

On other hand is used to create several subsets of data from training sample chosen randomly with replacement. Each collectionof subset data is used to train their decision trees. The reason why wehighlighted thesetwo techniques here is that a random forest is a bagging technique [2].

Random Forest Regression helps to improve the accuracy and it also decreases the over fitting in decision Trees. There is no need to normalize the data because it uses Rule based approach. This model is flexible for both classification and regression problems. It will work for both categorical and continuous values very well. It requires more power and resources it builds numerous trees to combine their output. Random Forest takes more time for training because it have to take combine number of trees. Due ensemble of decision trees, it fails to determine the significance of variables.

Decision Tree Regression is also supervised learning which is one of the most commonly used approaches [4] [5]. "Decision Tree can be used for both regression as well as the classification problems [7] [8]. Decision Tree is used to predict the target variable values in a non-linear fashion which means the best fit line is not going to be a straight line [14] [15]. Let's see some of the terms related to this model [20]." Decision Tree can be used for both regression as well as for the classification problems.

Decision Tree is very easy to understand and visualize. In this model, less data preparation was required. It can handle both numerical and categorical variables. Over fitting is one of the main disadvantages of this decision tree regression model. This problem can be solved by pruning. There is no guarantee to return the global optimal decision tree. It will not fit for continuous variables because the decision tree will loss the information when it was categorizes the variables into different categories.

IV. Proposed Model

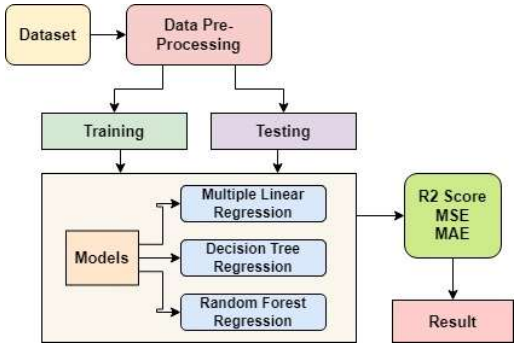


Figure 1. Proposed Model

Consider Rainfall and crop data as the input parameters and Crop yield as output Parameter.

1. Step 1: Now transform the taken raw data into information to obtain the results [14]
2. Step 2: partitioning the data for the dataset
3. Step 3: Applying the Regression models on trained data
4. Step 4: Calculate the performance of a model by evaluating the Mean Absolute Error (MAE), R2 Score and Mean Square Error (MSE)
5. Step 5: Model with high R2 Score are considered to be the best model for Crop Yield Prediction

Evaluation Metrics

R-Squared:

"R-Squared is a statistical measure that represents the proportion of variance for a dependent variable that is explained by the independent variables in regression model. It is also known as Coefficient of Determination [8][17]."

Formula for R-Squared is:

$$R^2 = 1 - \frac{UnexplainedVariation}{TotalVariation}$$

The formula will tell us the relationship there in between two variables. R-Squared is the square of correlation coefficient which is represented as 'R' (hence, it termed as R-Squared). The R-Squared is the value between o and 1 (i.e., 0 < = R-Squared < = 1).

If the R-Squared value is equal to zero (i.e., R-Squared value = 0) ,then there is no correlation and likewise, if the R-Squared value is equal to one (i.e., R-Squared value = 1), then it is a perfect correlation. The Closer the value of R2 Square to "1",the better is the model fitted and we can find the best model based on that. The higher R-Squared value indicates that there is better goodness of fit for the observations and model. For example, an R-Square value of 70% indicates that 70% of data only fit for the regression model.

Mean Square Error (MSE) :

"Mean Square Error is defined as mean or an average of Square of the difference between actual value and the predicted value. In supervised Learning method, the dataset contains both dependent and independent variables. We will build the model using independent variables and predict the dependent variable"[17]. It is mainlyused to check how close the predicted value to the actual value. We are using this model evaluation methods for Regression models.If the got lower Mean Square error value it says that the model is a best fit model.General formula for Mean Square Error is,

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \tilde{\mathrm{y}}_\mathrm{i})^2$$

Mean Absolute Error (MAE):

"Mean Absolute Error is one of the evaluation metric used with the regression models. The mean absolute error value of a model with respect to a test set is mean of absolute values of prediction errors[9]. Every predicted error is difference between true value and predicted value. Less mean absolute error value of a model indicates that the model was best model"[17]. It can be calculated by,

$MAE = \sum_{i=1}^{n} |y_i - x_i|$
Where, y_i is predicted value, $x_i$is true value.

IV. Experimental Results

We designed this complete system by using python. There are many datasets for example: crop, location, state, season, Area, production, rainfall.

Figure 2 gives the correlation between area, production and rainfall. There is a positive correlation between area and production as well as production and rainfall.

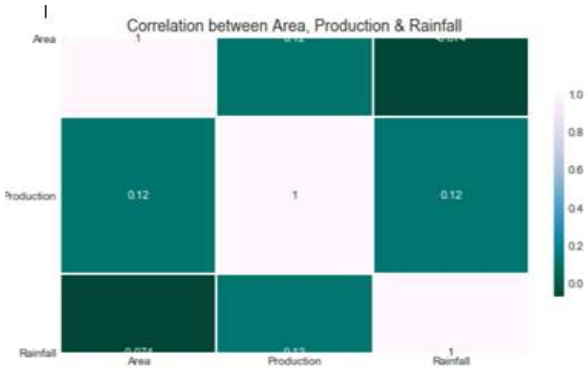


Figure 2. Correlation between attributes

Figure 3 gives which crop has the highest production among all.

```
Out[30]: Crop
         Rice                186350564530.0
         Wheat               148458427462.0
         Sugarcane            56527377710.0
         Maize                55366582627.0
         Potato                6147557322.0
         Cotton(lint)          5442732000.0
         Jute                  2375176869.0
         Masoor                1436769190.0
         Rapeseed &Mustard     1268339758.0
         Moong(Green Gram)      896555909.9
         Name: Production_x, dtype: float64
```

Figure 3. Comparison of the crops

Rice has the highest production 186350564530.0 in the dataset including the state in group by. Figure 4 gives which state is highest in the production along with the crop name.

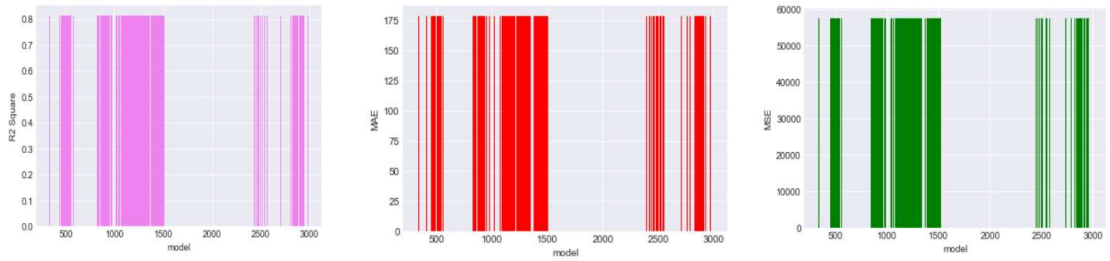
```
Out[29]: State   Crop
         Bihar   Rice          125611288530.0
         Punjab  Wheat          93607748000.0
                 Rice           60739276000.0
         Bihar   Wheat          54850679462.0
                 Maize          53603194627.0
                 Sugarcane      39304057710.0
         Punjab  Sugarcane      17223320000.0
         Bihar   Potato          6147557322.0
         Punjab  Cotton(lint)    5442732000.0
         Bihar   Jute            2375176869.0
         Name: Production_x, dtype: float64
```

Figure 4. Highest production of the state

Evaluation metric for Multiple Linear Regression:

```
R2 square: 0.8129076580669361
MAE:  178.35381062997376
MSE:  57500.51874258797
```

In Multiple linear regression the R2 score value is 0.81 and MAE value is 178.353, MSE value is 57500.518 (Figure 5).



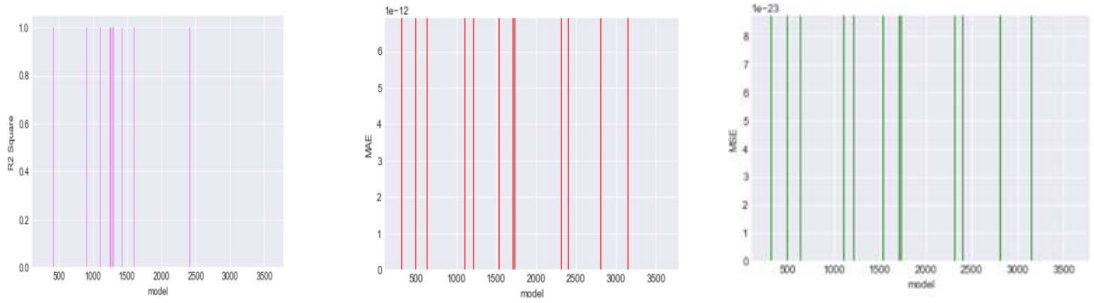
(a) (b) (c)

Figure 5. (a) R2 (b) MAE (c) MSE score for multiple linear regression

Evaluation metric for Decision Tree Regression:

```
R2 square: 1.0
MAE:  6.911750838921653e-12
MSE:  8.735288546485627e-23
```

In Decision Tree Regression the R2 score value is 1.0 and MAE value is 6.911, MSE value is 8.735 (Figure 6).



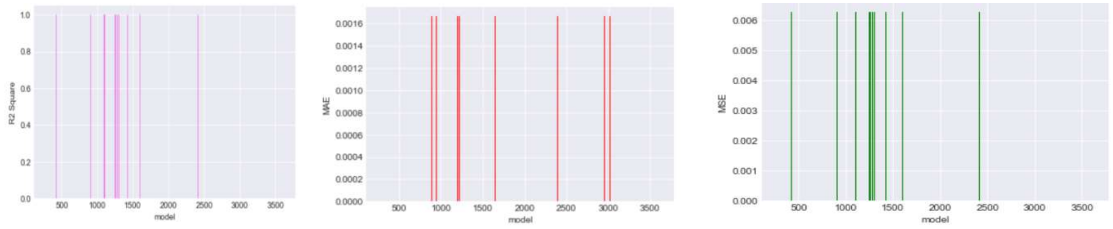
(a) (b) (c)

Figure 6. (a) R2 (b) MAE (c) MSE score for decision tree regression

Evaluation metric for random forest regression:

```
R2 square: 0.9999999795876571
MAE:  0.0016657775058753581
MSE:  0.006273481285519524
```

In Random Forest Regression the R2 score value is 0.99 and MAE value is 0.0016, MSE value is 0.0062 (Figure 7).



(a) (b) (c)

Figure 7. (a) R2 (b) MAE (c) MSE score for Random forest regression

V. Conclusion and Future Scope

This paper mainly focuses on finding the correlation between different attributes. This work mainly shows the comparison of different algorithms when we apply this on our dataset and it shows the R2 Score, Mean Square Error and Mean Absolute Error for each algorithm for training the dataset and it predicts that the crop named "Rice" in Bihar state has the highest production in our dataset and positive correlation exist in our dataset. After seeing the results, we can conclude that decision Tree Regression is the best fit model for predicting the crop yield prediction because it is has high R2 Score (i.e., 100%) and less Mean Square Error (MSE) and Mean Absolute Error (MAE) when compared to Multiple Linear Regression Model and Random Forest Regression Model.

This research work will enhanced to next Level. We will build a recommended system for farmer which will be more user friendly and they can make decisions on which season they can grow good crop. So they will get good profit. To automate the process by showing the prediction results in web application.

References

[1] Savla, Anshal, Nivedita Israni, Parul Dhawan, Alisha Mandholia, Himtanaya Bhadada, and Sanya Bhardwaj. "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture." In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-7. IEEE, 2015.

[2] Shakoor, Md Tahmid, et al. "Agricultural production output prediction using supervised machine learning techniques." 2017 1st international conference on next generation computing applications (Next Comp). IEEE, 2017.

[3] Doshi, Zeel, et al. "Agro Consultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms." 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, 2018.

[4] Bondre, Devdatta A., and Santosh Mahagaonkar. "Prediction of Crop Yield and Fertilizer Recommendation Using Machine Learning Algorithms." International Journal of Engineering Applied Sciences and Technology 4.5 (2019): 371-376.

[5] Nishant, Potnuru Sai, et al. "Crop Yield Prediction based on Indian Agriculture using Machine Learning." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.

[6] Dahikar, Snehal S., and Sandeep V. Rode. "Agricultural crop yield prediction using artificial neural network approach." International journal of innovative research in electrical, electronics, instrumentation and control engineering 2.1 (2014).

[7] Kumar, Rakesh, et al. "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). IEEE, 2015.

[8] Vijaya baskar, P. S., R. Sreemathi, and E. Keertanaa. "Crop prediction using predictive analytics." 2017 International Conference on Computation of Power, Energy Information and Communication (ICCPEIC). IEEE, 2017.

[9] Chlingaryan, Anna, Salah Sukkarieh, and Brett Whelan. "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review." Computers and electronics in agriculture 151 (2018): 61-69.

[10] Akshatha, K. R., and K. S. Shreedhara. "Implementation of machine learning algorithms for crop recommendation using precision agriculture." International Journal of Research in Engineering,Science and Management (IJRESM) 1.6 (2018): 58-60.

[11] Nigam, Aruvansh, et al. "Crop yield prediction using machine learning algorithms." *2019 Fifth International Conference on Image Information Processing (ICIIP)*.

[12] Abbas, Farhat, et al. "Crop yield prediction through proximal sensing and machine learning algorithms." *Agronomy* 10.7 (2020): 1046.

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

[14] Gandge, Yogesh. "A study on various data mining techniques for crop yield prediction." *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*. IEEE, 2017.

[15] Saranya, C. P., et al. "A SURVEY ON CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS." (2020).

[16] Ricciardi, Vincent, et al. "An open-access dataset of crop production by farm size from agricultural censuses and surveys." *Data in brief* 19 (2018): 1970-1988.

[17] Van Klompenburg, Thomas, AyalewKassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review." *Computers and Electronics in Agriculture* 177 (2020): 105709.

[18] Crane-Droesch, Andrew. "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture." *Environmental Research Letters* 13.11 (2018): 114003.

[19] PS, Maya Gopal. "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms." *Applied Artificial Intelligence* 33.7 (2019): 621-642.

[20] Kumar, Y. Jeevan Nagendra, et al. "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector." *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2020.