Design and Implementation of Various Regression Models for Yield Prediction

Vishal Jain, Vaidhehi V

***Abstract**: Agriculture is the backbone of India. In order to support farmers in India, this research is focused on the design of various predictive models that are used to predict the yield value for a specific crop in Indian states. This research work considers Rice, Wheat, and Bajra crops in Tamil-Nadu, Rajasthan, Uttar Pradesh states respectively. The various regression models such as Linear, Multiple, C4.5 and Random Forest are considered in this work. R squared value is used to evaluate the performance of the regression models. The result of this work shows that Random Forest model is better in performance.*

***Keywords** : Predictive analysis, regression models, Linear regression, Multiple regression, Random forest and C4.5 algorithm.*

I. INTRODUCTION

The Indian agriculture sector was valued at INR 16,587 crores in 2018. The market is calculated to succeed by 2024, increasing at a CAGR of 10.8% throughout 2019-2024. The agriculture market constitutes one of the foremost necessary fields of the Indian economy. The Republic of India presently serves as the world's biggest producers of the many recent fruits and vegetables, major spices, elite fibrous crops like jute, many staples like millets and physic seed. Additionally, India is the second-largest producer of wheat and rice, the world's major food staples. Currently, India ranks inside the world's 5 largest producers of over the eightieth percent of agricultural elements, together with several cash crops like cotton and coffee [1].

The enhancement of agricultural growth is not a simple task to achieve, the total area of land in India is unevenly distributed on basis of certain variables like soil quality, soil type, temperature around nearby areas, total production area and yield of particular crop, rainfall, MSP (Minimum Support Price) and many others. Understanding each factor and its effects on crops is a lot of work and collecting all the information accurately and providing it to the system for prediction is challenging and sophisticated on its own. Every region has completely different values for each variable which are mentioned above and because of its unbalanced distribution, it’s hard to get proper knowledge and data about a particular crop type. Basically finding the right method for the right problem.The focus of this paper is to do prediction on crop data with different prediction algorithms and get the approximately accurate crop yield for the particular crop in a region and also comparing those results from each algorithm to find the best fitted predictive algorithm for crop data. In order to do that, the comparison of each R squared value from the algorithm and finding the algorithm with less error rate. That algorithm will be the best algorithm for the crop data.

R squared is a method which tells how much of a dependent variable is understood by the independent variable. With help of this method, we will get a value which is known as R squared value and each algorithm’s R squared value is going to be compared and analyzed which eventually will give the best algorithm amongst the algorithm have chosen. Algorithms which are going to be part of this study is, Simple Linear Regression, Multiple Linear Regression, Random Forest, C4.5

1. About Algorithms

The various regression models such as simple Linear regression, Multiple regression, C4.5 and Random Forest algorithms are discussed.

A. Simple linear regression

It is the simplest form of analysis in regression as it uses one independent variable and one dependent variable. The relationship between 2 variables is claimed to be settled if one variable is often accurately expressed by the opposite by straight line.

Equation for Simple Linear Regression:

$$y = \beta 0 + \beta 1 * x + \varepsilon$$

The $\beta 0$ is the Intercept. The $\beta 1$ is the Slope. This y is the dependent variable and x is the dependent variable, ε which is the error term.

The goal is to find estimated values for $\beta 0$ and $\beta 1$ which would provide the ‘best fit’ for the data points. A line that minimizes the differences between actual and predicted values of the dependent variable ‘y’, each of which is given by, for any parameter values of $\beta 0$ and $\beta 1$.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author

Vishal Jain*, Computer Science Department, Christ (Deemed to be University), Bangalore, India. Email: jainvishal1896@gmail.com

Vaidhehi V., Computer Science Department, Christ (Deemed to be University), Bangalore, India. Email: Vaidhehi.v@christuniversity.in

B. Multiple Regression

Multiple regression tries to model the link amidst two or more independent variables and a response by fitting expression to determined knowledge; it is also called by the name "Multivariate Regression". The steps to perform Multiple Linear Regression are nearly just like that of Simple regression because it's the extension of the latter. The distinction lies within the analysis, will use it to search out that the issue has the best impact on the expected output and currently different variables relate to each other. This type of regression technique has many to one relationship, which allows one dependent variable and many independent variables.

Multiple Linear Regression Model:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots \beta pxp + \varepsilon$$

Modal has 'y' as our dependent variable and 'x1', 'x2' is our independent variable till 'xp'. $\beta 0$ is the intercept and, $\beta 1$, $\beta 2$ are the coefficients till βp. ε is the error term. In this, each coefficient is explained as the estimated change in 'y' corresponding to a one-unit change in a variable, when all other variables are held constant. As before doing the Multiple Linear Regression, we should have to follow some steps as our prep-work, that had generated the list of independent variables and dependent variables. Data is collected for related variables from different sources. As already mentioned before that, the Multiple Linear techniques are the same as Simple Linear Regression except the former has many independent variables and one dependent variable. So, as to conduct Simple Linear Regression for each Independent variable with dependent variable pair keeping other independent variables as constant. The tool has been used which makes this process automatic, faster and efficient. In the end, we will use the best fitting model to make predictions about the dependent variable.

C. Random Forest

Random Decision forest is a method that operates by constructing multiple Decision trees during the training phase, the majority of the trees is chosen by the random forest as the final decision. Basically, many decision trees provide the output of its own and then random forest checks for the majority of the output which is similar and selects the output as a final decision. Each decision tree has a different condition to narrow the dataset because right now the data is unpredictable and also there is a lot of randomness of data, which derives high Entropy.

$$E = \sum_{i=1}^{c} -p_i \log_2 p_i$$

So, to reduce the entropy each decision tree has different conditions and the dataset is divided and classified according to conditions of each branch in a decision tree. As the lower Entropy shows information gain, which is the measure of the decrease in entropy after the dataset is split. Measuring the information gain by finding the difference between higher entropy decision dataset and lower entropy decision dataset. Weka does all these processes in an efficient way.

D. *C4.5*

This algorithm is an improvised version of ID3 but has the same functionality with some extra features. ID3 has the problem of the Over-fitting of the model, but C4.5 resolved this issue. C4.5 generates a tree with the same as ID3 but prunes the unwanted or less effective branches, by which the Over-fitting issue is resolved. By pruning the accuracy of the model gets increased. The attribute showing the highest gain ratio is selected to divide the decision tree. The algorithm removes the bias value of information gain when there are more outcomes of a single attribute.

$$Gain\ Ratio(S, A) = \frac{Gain\,(S, A)}{Split\ Info\ (S, A)}$$

$$Split\ Info(S, A) = -\sum_{v=1}^{n} \frac{|S_v|}{|S|}\ Entropy\,(S_v)$$

The gain ratio is the modification of Information Gain which basically reduces the bias value. It takes size and number of branches before choosing an attribute.

The thorough study has been done on related papers which are mentioned in the 'Section 2'. The details about the attributes which are selected for the study has defined in 'Section 3'. The result of the implemention is mentioned in the 'Section 4' of the paper.

II. LITERATURE REVIEW

Veenadhari [2] et.al considers districts of Madhya Pradesh for computing crop productivity using environmental condition parameters. A sign of the relative influence of various climate parameters on the crop yield, different agro-input parameters accountable for crop yield is discussed in [2]. Sellam [3] et.al shows a relationship between AR (Annual Index) to FPI (Food Price Index) using multivariate analysis.

Aditya Shastry [4] et.al designed various regression models for the crops such as Wheat, Maize and Rice. The various predictive models in [4] are evaluated using mean square, R-squared values. Fernandes [5] designed a yield prediction model for sugarcane in brazil using decision tree model by considering various parametrs such as 10-day periods of SPOT-Vegetation(Satellite Pour I'Observation de la Terre), NDVI-pictures(Normalized difference vegetation index) and ECMWF(European Centre for Medium-Range Weather Forecasts) meteoric information. Wu et.al [6] implemented decision tree classifier to classify agriculture information. Kokilavani et.al [7] implemented the predictive model to spot the potential districts for the cultivation of rice, maize, and groundnut in Chennai by using attributes such as area of cultivation land, production and productivity of crops, Relative spread Index (RSI) and Relative Yield Index (RYI).

Ramesh [8] studied the yield prediction for crops using Density-based clustering methods. Shridhar [9] implemented the crop prediction system using IoT and Machine Learning techniques. SVM classifier [9] is used to predict the different changes in weather. K means [9] clustering method is employed to classify the soil and plants.

Prajakta [10] designed an Artificial neural network for crop prediction. In [11] states that the speedy advances in sensing technologies and machine learning techniques can offer efficient and comprehensive solutions for raised crop and setting state estimation and decision making.

III. OVERVIEW OF DATA

The data used for this research work is obtained from districts of Rajasthan, Tamil Nadu, and Uttar Pradesh in India. The preliminary data collection is carried out for some regions in all 3 states in India. And the crops which are going to be evaluated are staple crops of the particular states. The crops which are considered in the study are Wheat, Rice, and Bajra.

The data are taken in seven input variables. The variables are ‘Area’, ‘Production, ‘Rainfall’, ‘Temperature’, ‘Minimum Support Price’, ‘Crop Type’, ‘State’. The attribute ‘Area’ specifies the area where the crop is produced in Hectares. Attribute ‘Production’ specifies the production of the crop in Metric Tons. Attribute ‘Rainfall’ specifies average rainfall in Millimeter. Attribute ‘Temperature’ specifies the average temperature in Celsius. Attribute ‘Minimum Support Price’ specifies the price in Rupees. Attribute ‘Crop type’ specifies the crop for which all data is collected which is Wheat, Rice, and Bajra. Attribute ‘State’ specifies the states of India in which the particular crop produced in the specific year. Attribute ‘Yield’ specifies crop measurement per unit area of land cultivation.

Table 1: Instances and attribute details.

Sl.no	Particulars	Count
1.	Instances (Balanced dataset of 500 instances of each crop)	1500 (500x3)
2.	Attributes	7

IV. METHODOLOGY

The Prediction analysis is a practice of extricating information from existing data sets in order to determine patterns and predict future outcomes and trends.

The process consists of data collection, data has to be collected which shows some relevance to the domain, different data types can be used such as ‘Nominal’, ‘Discrete’, ‘Continuous’, ‘Spatial’, and ‘Time series data’. After the collection of data from many sources, it is said to be called ‘Raw data’, which is not at all feasible for further analysis.

So preprocessing is done to make it suitable. It includes checking of missing data, categorical data, standardization of data, and data splitting. Random Forest and C4.5 which are used in the study are fully capable of handling missing data, which gives the pure advantage to the researchers.

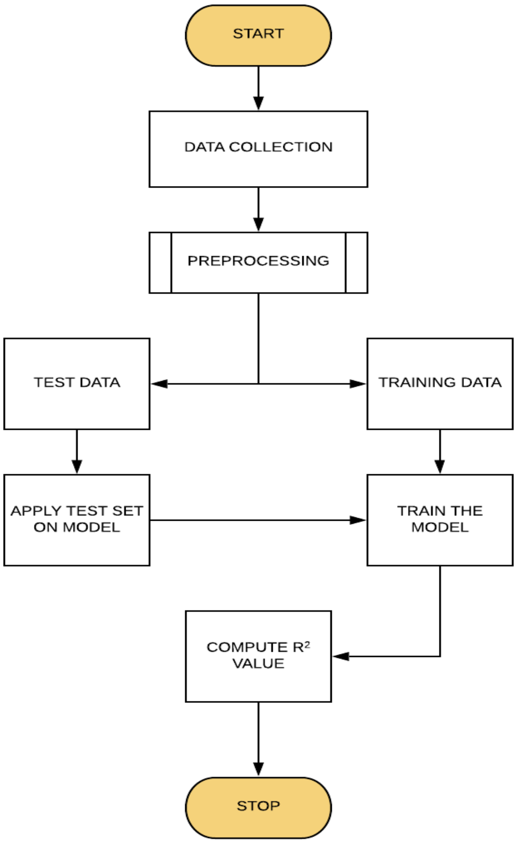

Fig 1: Basic steps of prediction analysis.

Data splitting into training and testing sets will be the final step for the Preprocessing. Here, the Pareto Principle for the data splitting which is 80/20 rule. And the crop data is divided into 80:20 ratios. Each regression model has trained afterward on 80% of the data; the trained model is tested using testing data.

Each time model trained and tested to find out the R squared value of the model, which is the error rate of the model. Measures have to be taken to get the less error rate, such as feature transformation, feature selection, algorithm tuning, and Ensemble methods (Bagging and Boosting). In the end, all the algorithms R squared values will be compared and which has less error rate will be the best-fitted algorithm.

Steps involved in the research:

1. Data collection – There are different ways of data collection; field research and collection of data, data from databanks and repositories, and auto generation. In the study, the collection of data is taken from databanks and repositories, where the ranges of each crop details is obtained and finally auto generated within that range. The final data with instance count mentioned in table 1 is used in this study.
2. Preprocessing – Attributes or feature selection is done by checking closely with the correlation risk factors. Dividing the data into two sets which is training set and testing set with the ratio of 80:20, according to Pareto principle.

3. Train the Model – Model training is done by using Weka, a tool which has algorithms and process them in the efficient way. Each algorithm which is used in the study has been trained on 1200 instances.
4. Apply Test set on the model – As the model is trained, now to test the model by the same dataset which is segregated before into training and testing by 80:20 ratios. The testing is done with 20 % of the test data set.
5. Computing the R^2 value – Each algorithm is trained, tested and an error rate is find out. In the study, the computation of R squared value shows the error rate for each algorithm. "Table 2" consists of R squared value for all the algorithms, more the R squared value means less error rate.

V. RESULT AND DISCUSSION

According to the dataset used in the study, which has 1500 instances, 7 independent variables and a single dependent variable which is 'yield'. The Evaluation Criteria for the algorithms would be R Squared value. This is a statistical measure that indicates how much of a dependent variable is understood by the independent variable(s) in a model.

Linear regression and Multi-Linear regression serves the same purpose of predicting the dependent variable, although the difference is the number of attributes or the independent variable which will be accountable for the variation in the dependent variable. According to the R2 value of all the algorithms, the analysis from this is, the best algorithm is the one who can handle the missing values properly and the one with greater R2 value will be the best out of all the algorithms.

Table 2: R squared value for all algorithms.

Sr.No.	Algorithms	R squared value
1	Simple Linear Regression	0.0032
2	Multiple-Linear Regression	0.0049
3	Random Forest	0.5232
4	C4.5	0.5025

The multi-linear regression model has greater R2 value, which serves the best result after training and testing on the dataset between the former and latter.

C4.5 and Random forest is the decision tree classifier and regression modal, C4.5 is the upgraded version of ID3, it handles missing values better than ID3 and it is way faster and create a single path decision tree. The random forest creates a tree for each node, which is much better than C4.5 but slower than C4.5. The R2 value for the Random Forest is better than C4.5. As mentioned earlier Random forest makes a decision according to the majority out of all decision trees and C4.5 is better at pruning.

The selected attributes were 7, the sensitivity and specificity by the same variables are also evaluated and checked. Reducing the number of attributes leads to less R squared value and increasing attributes created correlation risk factors. So, the study has been stopped at this point and the table above is the point where the study gets the highest R Squared value of the model using the Crop dataset.

VI. CONCLUSION

Machine learning is achieving its popularity in almost all operations of the real world. One of the Machine Learning techniques i.e., Regression and prediction analysis is an interesting topic to the researchers as it predicts the value based on given data and tries to find a pattern in the data. This method extracts knowledge from data by using statistical tools and understands the relationship between different attributes.

This comparative study between some predictive algorithms with crop dataset helps to understand the different attributes which affect the crop yield and also which directly alter the yield variable, better understanding of different algorithms such as Simple Linear, Multi Linear, Random Forest and C4.5 implementation over crop yield dataset. The experimental result shows Random Forest is offering the best result in the case of our dataset amongst all. It is showing less error rate, which is taken as a high R squared value. This study can be expanded by using different prediction models which can be evaluated with different metrics. The hybrid model can be designed by grouping up different predictive algorithms. The number of crops and the number of states can be considered for designing the models.

REFERENCES

1. Indian Farming Market Share, Size, Growth, Demand and Forecast Till 2024: IMARC Group, Retrieved June 29, 2019
2. From http://www.emailwire.com/release.
3. Veenadhari, S., Misra, B., & Singh, C. D. (2014, January). Machine learning approach for forecasting crop yield based on climatic parameters. In 2014 International Conference on Computer Communication and Informatics (pp. 1-5). IEEE.
4. Sellam, V., & Poovammal, E. (2016). Prediction of crop yield using regression analysis. Indian Journal of Science and Technology, 9(38), 1-5.
5. Aditya Shastry, H.A. Sanjay and E. Bhanusree, 2017. Prediction of Crop Yield Using Regression Techniques. International Journal of Soft Computing, 12: 96-102.
6. Fernandes, J. L., Rocha, J. V., & Lamparelli, R. A. C. (2011). Sugarcane yield estimates using time series analysis of spot vegetation images. Scientia Agricola, 68(2), 139-146.
7. Wu, J., Olesnikova, A., Song, C. H., & Lee, W. D. (2009, January). The development and application of decision tree for agriculture data. In 2009 Second International Symposium on Intelligent Information Technology and Security Informatics (pp. 16-20). IEEE.
8. Kokilavani, S., & Geethalakshmi, V. (2013). Identification of efficient cropping zone for rice, maize and groundnut in Tamil Nadu. Indian Journal of Science and Technology, 6(10), 5298-5301.
9. Ramesh, D., & Vardhan, B. V. (2015). Analysis of crop yield prediction using data mining techniques. International Journal of research in engineering and technology, 4(1), 47-473.
10. Shridhar Mhaiskar , Chinmay Patil , Piyush Wadhai , Aniket Patil , Vaishali Deshmukh, "A Survey on Predicting Suitable Crops for Cultivation Using IoT", International Journal of Innovative Research in Computer and Communication Engineering.
11. Mrs.Prajakta Prashant Bhangale[1], Prof. Yogesh S. Patil[2], Prof. Dinesh D. Patil[3].Improved Crop Yield prediction Using Neural Network.
12. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and electronics in agriculture, 151, 61-69.

AUTHORS PROFILE

Mr. Vishal Jain received his BCA degree from Bhupal Noble's College, MLSU(MohanLal Sukhadia) University, Rajasthan. He is currently pursuing his MCA degree in CHRIST (Deemed to be University), Bengaluru, Karnataka. He is a Front end developer and his areas of interest include Mobile Application and Machine Learning.

Prof. Vaidhehi V received her MSc degree from Bharathidasan University, TamilNadu and MPhil degree from Kamaraj University, TamilNadu. She is pursuing her Ph.D. at Jain University, Bengaluru, Karnataka. Presently working as Associate Professor in CHRIST (Deemed to be University), Bengaluru, Karnataka. Her areas of interest include Machine Learning and Recommender Systems.