International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-10, August 2020

Crop Yield Prediction using Regression Model

Shikha Ujjainia, Pratima Gautam, S. Veenadhari

Abstract*: This research is done to find out the production dependability of crop with various physical circumstances. The prediction can also be done of a crop yield by using the model of regression and it is mainly discussed in this paper. Machine learning is an emerging research area in Agriculture, particularly in crop yield analysis and prediction. There are some complex data which are tough to decode or find by everyone, the strategies of machine learning can be used in this scenario and automatically the valuable underlining pattern can be accessed. Various complex decision-making activities can be performed when the feature of machine learning will enable the knowledge and patterns which are unseen about any problem. The future events can also be predicted. In the growing season as possible, a farmer is focused on conceptualizing how much yield they except. Like many other regions, the amount of agricultural data is increasing at the daily source. This paper aims to predict crop yield on the collected agricultural dataset. The regression analysis model is used to test the accuracy and effective predictions of the rice crop yield in India. Linear regression is used to establish a relationship between various environmental variables like temperature, rainfall, etc and the crop yield. It is important to measure the possible production of rate of crop and the farmers will be benefitted by the result of this prediction. As financial impact is attached of the farmers with the yield production, the research will support them to avoid any loss. The accuracy of the prediction through regression model is also observed in this research paper.

Keywords: Machine learning, Regression model, Linear regression, Yield prediction.

I. INTRODUCTION

Agriculture and its related sectors are undoubtedly the largest livelihood provider in India, especially in rural areas. The main challenge in the field of agriculture is to raise the grain productivity per unit of land. By the use of emerging technologies, we can help the farmer to predict the crop yield or forecast the production of the crop for the next year with the change of various agro-climatic conditions. The productivity of the crop is majorly influenced by weather conditions. Therefore, accurate yield prediction is a major problem that must be resolved. Internet of things (IoT) makes sense when we want to collect real-time data, various sensors like temperature sensors, humidity sensors, soil moisture sensors, location sensors, etc are inserted into the field to collect the data. This data further used in various predictions like crop yield prediction, soil fertility measures, insect detection, etc by using machine learning algorithms. The core developing element of machine learning are technologies of big-data and high-performance computing which creates opportunities for the field of data-intensive science that is used in the sector of multi-disciplinary agro-technology. Among various definitions, ML is described as the logical field that enables machines to learn without being carefully modified [1] [2]. It alludes to the capacity of a machine to anticipate the result without being expressly customized. There are an enormous number of decisions for ML apparatuses. An application master needs to settle on a reasonable decision on a particular ML technique to send for his/her particular issue. We advocate that the specialist should down select from the plenty of ML decisions dependent on the sort and measure of accessible information and issue detailing [3]. From the perspective of crop yield prediction, we identify the relationship between environmental variables and crop yield production. Furthermore, the data preprocessing step will play a crucial role in successful decision-making. AI strategies which are broadly utilized in forecast procedure are boosting methods (for example RGF, GBDT, and Add support), relapse tree (for example ID3, C4.5, and M5-prime relapse tree), straight relapse, arbitrary timberland, bolster vector machine, k-closest neighbors and fake neural system. Among all these expectation strategies boosting procedures (for example GBDT, RGF) is as yet immaculate in crop yield forecast [4]. Through this research, the linear regression technique is studied and the corresponding analysis is presented.

Fig. 1.Machine leaning process

If you want to do prediction or forecasting using machine learning algorithms, then you must follow the basic steps presented in fig.1. In the data collection step, there is the various source by which we can get data. Next, the data is bifurcated by preprocessing methods, in which categorical and null values are handled. Appropriate algorithms is to be selected according to build the model. The data is further divided into training and testing modules to test the accuracy of our model. And the last, results are visualized through flow chat, CSV format, excel format, or other visualization tools.

Revised Manuscript Received on August 30, 2020.
* Correspondence Author
Shikha Ujjainia*, Department of Computer Science and Application, Rabindranath Tagore University, Bhopal, India. E-mail: shikhaujjainia90@gmail.com
Pratima Gautam, Dean (CSIT), Rabindranath Tagore University, Bhopal, India. E-mail: pratima_shkl@yahoo.com
S. Veenadhari, Associate Professor (CSE), Rabindranath Tagore University, Bhopal, India. E-mail: veenadhari1@gmail.com

II. OBJECTIVES

The main objective of this research is to predict the crop yield with the help of a linear regression process. This will help to know the future situation of the production level of the yield and give ideas to the farmers to avoid the loss. Some of the specific objectives are:

i. To know the relation in between the production level of the yield and temperature. The various effects of the temperature and how it controls the crop yield will be determined.
ii. To find the effect of rain in the process of farming and how it is affecting the crop yield prediction as well as production.
iii. The area of the yield and the relation of this with the production of the crops will be found in this research.
iv. To determine the effect of ground-level water in production.
v. To find the production model's efficiency and how it can help the process of farming. The differences in the value which are predicted and actually seen in the real life are examined.

III. LITRATURE REVIEW

In the year of 2011, there was a study by Zaefizadeh with the group of some other researchers about the various ways of data mining and discussed how the prediction for crop yield is done by the technologies of data mining. In Ardabil, forty genotypes were planted. To do the prediction of grain yield, the application of artificial neural networks and multiple linear regression was done [5]. Fifteen neurons along with one hidden layer were implemented in artificial neural networks (ANN) [6]. Through the findings of Zaefizadeh, it was found that multiple linear regression outperformed artificial neural networks.

The research done by Sanchez in the year of 2014 showed the correlation between the linear and nonlinear strategies to do the prediction of crop yield. By performing a complete algorithm and the percentage split validation, a comparison was done and the most useful property subset for each strategy was found. The performance was found and calculated through test datasets which is consists of an unseen database. The data-driven process of prediction of the crop yield is mostly recognized and various methods were assessed by Sanchez in his research. The research has a various field which can be extended to for the huge number of crop datasets and techniques. Zhang in his research in the year 2010 showed the model of linear regression. For the prediction of crop yield, the utilization of the estimation process of the ordinary least square is used. According to this research, apart from the temperature, the precipitation contributed to the yield of corn. In the year of 2009, another research was done by Zaw and Naing and discussed the model of Polynomial regression model about the prediction of crop yield [7]. This research was done based on Myanmar and to predict the rainfall of that region.

IV. METHOD USED

The objective of this research is to show the impact of weather parameters and soil parameters on the yield production to improve the crop yields, which will benefit the farmers. The linear regression model is formed in the python.

A. Data collection

For the study, the statistical information is collected from Kaggle.com. The dataset consisting of historical data to be taken for rice.

The variety of attributes are regarded as following:

- Area (In Hectare)
- Temperature (Degree Celcius)
- Rainfall (mm)
- Groundwater level (m)
- Soil Ph
- Potassium (kg/Hectare)
- Magnesium (kg/Hectare)
- Sodium (kg/Hectare)

B. Data preprocessing and feature extraction

The modifications applied before feeding it to the algorithm are referred to by preprocessing. Data preprocessing is a technique used to convert data into a data collection that is fresh. Additionally, data is gathered from other sources it's collected in a format that isn't possible for analysis. It's required to data preprocessing for achieving outcomes from the applied model in machine-learning.

Feature Extraction is a logically wide procedure where one attempts to build up a change of the information space onto the low dimensional subspace that jam a large portion of the significant data [8] [9]. Highlight extraction and determination techniques are utilized detached or in blend to improve execution, for example, evaluated precision, perception, and intelligibility of scholarly information [10]. As a rule, highlights can be sorted as: applicable, immaterial, or repetitive. In the component choice procedure, a subset from accessible highlights information is chosen for the procedure of the learning calculation. The best subset is the one with minimal number of measurements that most add to learning precision [11][9].

C. Regression Analysis

Regression analysis is a type of predictive modeling procedure that analyzes the connection between a dependent or target variable and independent or predictor variable (s).

It includes several models, such as linear, multiple linear, and non-linear regression. The most common models are simple linear regression and multiple linear regression. Non-linear regression analysis is commonly used for more complicated data set in which the dependent and independent variables show a nonlinear relationship.

D. Linear Regression

Linear regression is examined as a procedure that is utilized to break down a reaction variable Y which changes with the estimation of the intercession variable X. A methodology of anticipating the estimation of a response variable from a given estimation of the explanatory variable is referred to as prediction. Here to find the relationship two variables, one is the dependent variable (Y) and the other one variable that is independent (X) with a best fit straight line is commonly called as regression line [12]. The regression equation is shown below,

$$Y = a + (b*X) + e$$

Where,

- Y – Dependent variable
- X – Independent variable
- a – Intercept
- b – Slope
- e – Residual (error)

Linear Regression is very sensitive to outliers. This can greatly affect the regression line and predicted values. One main reason to select the linear regression is that the parameters getting, are continuous in nature and linear regression work best in the continuous variables.

If the independent variable has more than one input parameter, multiple regression can be implemented. The numerical representation of multiple linear regression is :

$$Y = a + (b*X1) + (c*X2) + (d*X3) + e$$

Where,

- Y - Dependent variable
- X1, X2, X3 - Independent variables
- a - Intercept
- b, c, d – Slopes
- e – Residual (error)

E. Crop prediction using regression method

Considering weather data (temperature, rainfall), crop data as the input parameters, and crop yield production as output parameters.

Step 1: Collect the data. Now transform this raw data into information. If the raw data is not enough to work with model, it will be necessary to apply duly designed format data to the model, to obtain suitable results.

Step 2: Now divide your dataset into two groups i.e. training and testing dataset. The training dataset is the subset of your dataset which is used to train your model whereas test dataset used to test your trained model. The training set will have a maximum rate of information to train most instances to produce. About 70% of the samples are collected under the training set. Remaining test dataset uses the information to check how the model is performing.

Step 3: Apply the linear regression algorithm on a trained dataset.

Step 4: Calculate model performance by evaluating R2, RMSE (Root Mean Squared Error).

Step 5: Apply that trained model on the test dataset and again calculate R2 and RMSE to measure the performance of the model. Model with the high accuracy and R2 values and the low RMSE statistics values are considered to be the best model for the crop yield prediction.

As studied earlier, now following figure is shown, predicting rice crop production with various weather and crop parameters. It is tried to correlate the independent variables to yield production which are presented in the figure:

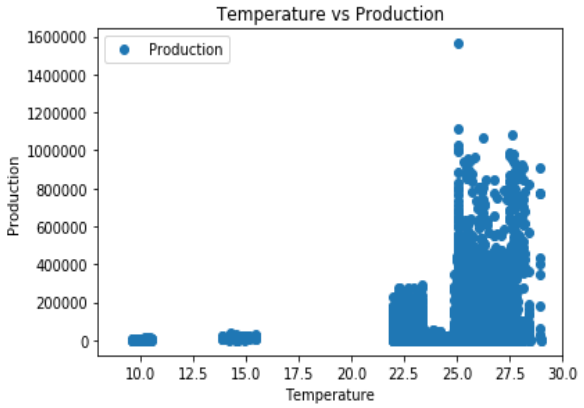

Fig. 2. Relation between temperature and yield production.

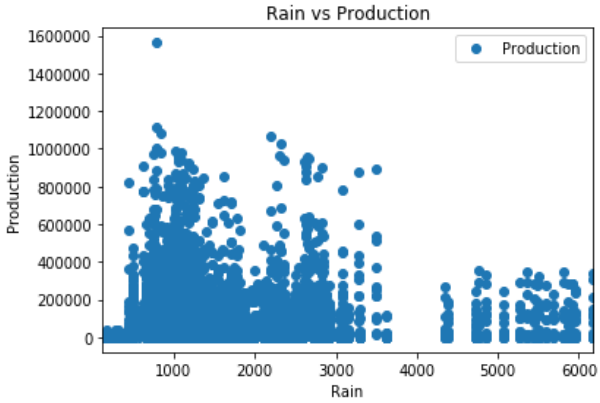

Fig. 3. The relation between rain and yield production.

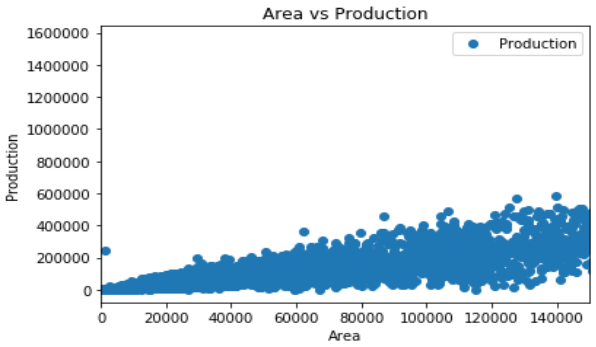

Fig. 4. Relation between area and yield production.

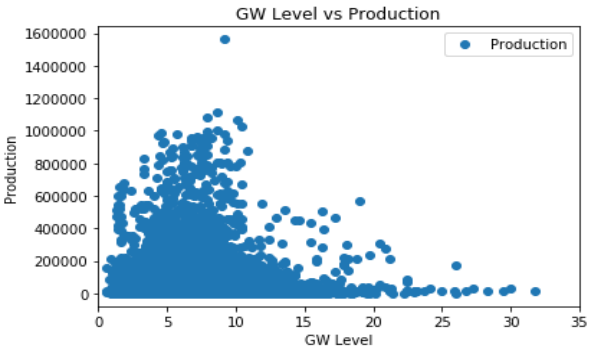

Fig. 5.The relation between ground-water level and yield production.

The ratio of 75% and 25% is fixed for training and testing dataset respectively.

This model is achieved R2 with 0.75 i.e. 75% accuracy. R2 is a square of the correlation between predicted target values 'y' and actual target values 'y' which falls in the ranges from 0 to 1. R2 of accurate 1 means the dependent variable exactly predicted the by from the independent variables, which never happens whereas if the value of R^2 becomes 0 means dependent variable cannot predict by from the independent variable. So, it always is good for the model to predict the value of R2 near to 1.

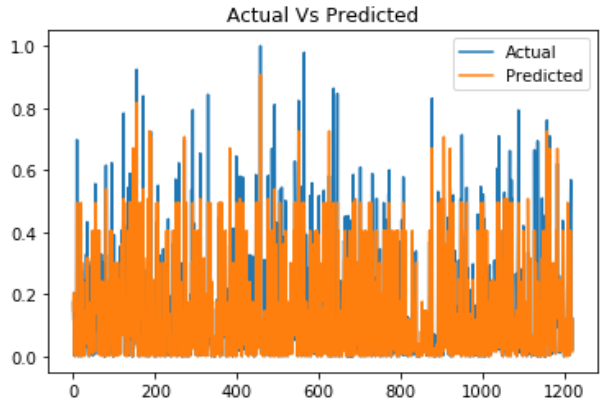

Fig. 6 Prediction of model between actual and predicted values.

F. Findings

A different important matter which are associated with the yield productions and drives the quality and quantity of crops are measured by this research. Driving factors along with the predicting methods were also discussed in this research. Some of the important findings of this research are:

i. The yield production level is dependent on the temperature of the atmosphere. The farming of crops is affected due to the fluctuation of temperature in different stages. The temperature has the ability to control the production level and how it affects crop production is also found through this research. The temperature was found one of the main driving factors in the process of crop yield prediction.

ii. Crop production is also depending on the amount of rain. Water is really essential for farming and the process of crop yield prediction is also needed to measure the contribution of it. The rain is important, but the amount of rain and the proper timing is there for a good level of crop production. This was measured in this research and how this is manipulating the process of yield production and the process of measuring this driving element was found.

iii. Yield production is attached to the area of yield where the crops are being planted. The production level direct proportionally changes with the area. Yield with more area has the ability to increase the production level of crops. This research gave clarity about this factor and it was also found that the production cost and maintenance efforts also increase along with the increase of area.

iv. The level of groundwater is also a driving factor in the process of yield production. If the level of groundwater goes down the crops will not grow properly, and the production level will be affected. This research shows that there should be a sufficient level of groundwater for the good production of crop yield.

v. The relation between the predicted production quantity of the crops with the driving factors and the actual production in the real-life scenario was determined. The predicted and the real-life numbers were not that much different to conclude the research as irrelevant rather the predictions made were so close to the real outcomes. The main finding of this research showed that this prediction procedure is very helpful for the process of farming and avoiding the loss of the farmers.

V. CONCLUSION

Food plays a vital role in survival for everyone. Farmers face lots of difficulties due to various unpredictable reasons. Hence to overcome the unpredictability of crop production or other agriculture-related problem we use some prediction models. The regression model is used as a prediction tool to predict crop yield.

Thus, in this work, linear regression analysis is used to establish a relationship between various independent parameters as explained above and their effects on rice yield, aim to increase crop productivity by using correct predictions with the model. Dependency of the production of crop with various parameters like temperature, rain, etc. is also determined in this paper and based on the result the prediction is done for crop yield. This research has shown the accuracy of the crop yield production while predicted values are compared with the actual production quantity. The problem that was faced by the farmers, the model of regression is able to give a permanent solution.

RECOMMENDATIONS

There are various parts in this research that needs further research to know more about crop yield prediction through the method of regression analysis. In future analysis can be done with other prediction methods. Support vector regression is one of the useful analyzation methods which can be applied in the same scenarios and various other information can be gathered for the objective of crop yield prediction. There are also some prediction techniques that can be used in future research like Fuzzy Logic, Neural Networks, etc. By using these methods, the yield of various crops can be predicted. In between the different predictor variables co-relation can be measured. As the variable is an important part of the process of prediction, that can also be found in future researches.

REFERENCES

1. A.L. Samuel, Some Studies in Machine Learning Using the Game of Checkers I, D. N. L. Levy (ed.). New York: Computer Games I, 1959.
2. K. Liakos, P. Busato, M. Dimitrios, S Pearson, and D Bochtis, "Machine Learning in Agriculture: A Review," Sensors, vol. 18, no. 8, pp. 1-29, August 2018.
3. A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, "Machine Learning for High-Throughput Stress Phenotyping in Plants," Trends in Plant Science, vol. 21, no.2, pp. 110-124, February 2016.
4. R. Kumar, M. P. Singh, P. Kumar and J. P. Singh, "Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique," International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy, and Materials (ICSTM), pp. 138-145, May 2015.

5. Olive, David J. "Multiple linear regression." In Linear Regression, pp. 17-83. Springer, Cham, 2017.
6. Van Gerven, M. and Bohte, S., Artificial neural networks as models of neural information processing. Frontiers in Computational Neuroscience, 11, p.114, 2017.
7. Schönbrodt, F., Testing fit patterns with polynomial regression models., 2016.
8. N. Chumerin and M. Van Hulle, "Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information," Proc. IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, 2006, pp. 343–348.
9. S. Khalid, T. Khalil and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," Science and Information Conference, London, pp. 372- 378, August 2014.
10. H. Motoda and H. Liu, "Feature selection, extraction and construction," Sixth PacificAsia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 67–72, 2002.
11. L. Ladha and T. Deepa, "Feature Selection Methods And Algorithms," International Journal on Computer Science and Engineering (IJCSE), vol. 3, no. 5, pp. 1787-1797, May 2011.
12. P. Surya, and I. L. Aroquiaraj, " Crop Yield Prediction In Agriculture Using Data Mining Predictive Analytic Techniques, " International Journal of Research and Analytical Reviews (IJRAR), vol. 5, no. 4, pp. 783-787, December 2018.

AUTHORS PROFILE

Shikha Ujjainia received B.Sc. degree in Computer Science from Career College, Bhopal, in 2011 and Post Graduation Degree in Master of Computer Application from Samrat Ashok Technological Institute (SATI), Vidisha, in 2014. She is currently undergoing Ph.D. programme with Rabindranath Tagore University, Bhopal, Madhya Pradesh. Her research work & interest includes machine learning, data analytics and data mining. As a research scholar, two research papers has been published in reputed journal. She is also actively involved towards online workshops, seminars, and conferences based on IoT, Machine Learning, and Big Data. She has also actively took part in similar programs organized by the university.

Dr. Pratima Gautam has PhD in Computer Application from Department of Computer Application, Maulana Azad National Institute & Technology, She is currently working as Professor in Department of Computer Science & Information Technology, Rabindranath Tagore University, Bhopal, India. With more than sixteen years of teaching experience. She has authored several research articles in International journals of repute. Besides having presented papers in several international/ national conferences. She has been invited as an expert to various national conferences as paper reviewer/ program technical committee member. She has delivered lectures in various institutions and has also participated in various training programs and attended several workshop. Her research interests include Data Mining, Machine Learning, Soft Computing and image processing. She is also affiliated with international societies like IEEE, , ACM digital library etc She has associated herself in guiding several under graduate and post graduate students in their projects and is currently providing Ph.D. supervision to six research scholars. She is also actively involved in Institutional activities like organizing Conferences/ Workshops etc.

Dr. S. Veenadhari completed her Doctoral programme from Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya in 2015, Master of Computer Applications from Nagarjuna University, Andhra Pradesh in 1998 and Master of Technology in Computer Science and Engineering from Makhan Lal Chaturvedi University, Bhopal in 2007. Over 15 years of academic and research experience with 45 research papers published in International and National reputed journals. Six students are pursuing their Doctoral programme and two students completed their doctoral degree under her supervision. Her research interests includes Machine leaning, Big data analytics and Cloud computing. Her expertise in agricultural data analytics through machine learning. She has published several book chapters, technical bulletins, project reports, working papers , training and teaching reference manuals. She has delivered number invited talks in many national platforms of repute. She has worked as Expert member in various committees in different institutions and member of different professional societies like IE, CSI, ISTE.