Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.9 No.1, 2020, pp.40-44
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2020.9.1.2152

Linear Regression Approach to Predict Crop Yield

R. Murugan, Flaize Sara Thomas, G. GeethaShree , S. Glory and A. Shilpa
Department of Information Science and Engineering, T John Institute of Technology, Karnataka, India
E-mail:muruganraam75@gmail.com

(Received 2 April 2020; Revised 13 April 2020; Accepted 2 May 2020; Available online 10 May 2020)

***Abstract* - The agriculture plays a very big and important role for the country's growth. The agriculture science system facing lots of problems from the environmental change. Machinelearning (ML) is the best approach to overcome the problems by building the good and effective solutions. Crop yield prediction include prediction of yield for the crop by analyzing the existing data by considering several parameters like weather, soil, water and temperature etc. This project addresses and defines the predicting yield of the crop based on the previous year's data using Linear Regression algorithm. The approach of this project is to solve the problem of cost loss. Real agricultural data is used for making the models and the models tested with the samples. The prediction model will help the end users (farmers) to predict the crop yield before cultivation of the crop onto the agricultural field. To predict the accurate results Linear Regression machine algorithm is used. The presence of large dataset will help to improve the decision making model.**

***Keywords*: Regression algorithm, Machine Learning, Crop Yield Prediction.**

I. INTRODUCTION

During 18th century, Agriculture was practiced as the main source of income in our pretty world. Where in India, once the trade was made using the grown crops spices, pulses etc.., which were treated as a source of exchange of goods and variety of grains and pulses despite of money under our Barter System which was practiced throughout. Due to global warming and pollution factors, there is limited or much decrease in the quality and production of crops. The scarcity of water, increase in temperature makes it impossible to give an optimal production in the crops. The tropical climatic conditions suit best for various types of crops to grow in India. India being the major producers of many crops such as cotton, spices and so much other important crops. There is a lot of impact on the climatic conditions and as well as the parameters such as soil, temperature, humidity. Such a valuable agriculture profession is going down in our India. Due to the factors like natural calamities (i.e., drought, flood etc.,) that affected our agricultural lands. despite of this some of the weather parameters (like rainfall, temperature, humidity) and soil parameters (like Nitrogen, Phosphorous, Potassium) also affect the growth of the crop.The tropical climatic conditions suit best for various types of crops to grow in India. India being the major producers of many crops such as cotton, spices and so much other important crops. There is a lot of impact on the climatic conditions and as well as the parameters such as soil, temperature, humidity. Each individual crop has its own requirement amount of restrictions with respect to the amount of temperature, rainfall, phosphorous, nitrogen, potassium etc.., it should consume to give the best yield in return. Example, soy bean crop, it requires 15% moisture and consumes 11kg P/ha with temperature above 13º C to provide its best yield. In such condition if a soy bean is grown in an area with high moisture and no phosphorous the crop dies.This is the same problem which is faced by our Indian farmers where most of them don't even have knowledge of the new crops which can give better yield with high profit if grown in that particular area. Instead, either they grow the different crop which doesn't even needs that amount of a parameter to grow or cultivate a crop which needs high parameter in a low parameterized area. Hence change of weather parameters and soil parameters pushes our farmers under unfortunate loss. While India is growing in its technological developments, bringing the aid of technology in the field of agriculture well provide farmers a great yield by providing a profitable production in crops. By using machine learning algorithms, it is proven to predict the kind of crop that can be grown in the particular area by inputting the climatic and soil parameters. In our proposed system we are using one of the machine learning technique which is multiple linear regression algorithm. Where the location of the crop is taken as input. From the location, the static data of soil parameters such as Nitrogen, Phosphorous and Potassium is obtained. And the weather parameters expected in current year is obtained from weather department. Those static datasets of the crop related to its production and demands of various crops taken from different websites of our government. Then the multiple linear regression algorithm and identifies the pattern among data and then process it as per input conditions. This will result in the best feasible crops according to the locations soil and weather conditions. Through this prediction, the farmers will be able to grow the crops based on the favorable conditions Thus, this system will only require the location of the user and it will suggest number of profitable crops providing a choice to the farmer about which crop to cultivate with the import and export profit details of past years. As past year production is also taken into account, the prediction will be more accurate.

II. LITERATURE SURVEY

S.Nagini, Rajinikanth, B.V.Kiranmayee (2016) states that agriculture yield prediction is one of the hardest job in agriculture field. The agricultural yield depends on various parameters like rainfall, soil moisture,water,nitrogen, surface temperature, etc.Since earlier the agriculture yield prediction was not done accurately after implementing many techniques also and farmers used to face lot of difficulties. After doing so much of research they developed predictive modeling. In this article these people had concentrated on two states that is Andhra Pradesh and Telangana to find the effective prediction or the forecast of the agriculture yield for various crops. They constructed various predictive models and they also collected information regarding explorative data analysis and various regression models were also used. Regression analysis is one of the predictive data modeling. Agricultural yield prediction is measured as production-in-Ton and there exists a correlation between area-in-hectares and production-in-tons which in turn means that as the area of land increases the agricultural production will also increase.

S.Veenadhari,BharatMishra, CD Singh (2011) says that soybean crop productivity can be measured accurately using decision tree algorithm. The chief crop of Madhya Pradesh is soybean. For 2 decades there were not able to get the good yield even if they increase the land for cultivation. And now by using decision tree, climatic factors and the major parameter for the soybean crop is the consideration of relative humidity which is done using decision tree and considering previous year’s data of crop productivity. By considering all these parameters they were able to predict accurately the yield of soybean crop. Decision tree tells us that there exists a correlation between climatic factors and soybean crop productivity. But the drawback is that this decision tree algorithm can be used only for soybean crop which is grown in MadhyaPradesh.This algorithm can’t be used for predicting any other crops.

Ms Shreya V. Bhosale *et.al* (2018) describes that there are different techniques to predict the crop yield where those techniques use K-means Clustering, Apriori, Naïve Bayes algorithm and the interrelation among these algorithms to predict the crop yield with Big data Analytics in our Indian Agriculture. In this paper, it is said that Naïve Bayes algorithm provides the probabilities of crop yield percent that is grown in that area. And says that the result will be crop name suggestions which can be grown in that area in accordance with the rainfall as well as number of acres of land of a farmer and gives average production of the crop per acre.

S. Bhanumathi *et.al* (2019)describes that in a populated country like India the climatic changes are common so oneneeds to secure the resources of food. Here the data mining technique is used for predicting crop yield by analyzing the previous year’s crop data.

The data of crop is first pre-processed in this paper, later the back propagation and random forest algorithm is applied for the data and the results of both the algorithm are compared to find the errors , simultaneously the back propagation algorithm is applied for fertilizer data .finally predicted yield and amount of fertilizer is displayed as result.

Yogesh Gandge,Sandhya (2017) suggests data mining techniques approach to provide more accurate results to predict the crop yield with smaller data sets. It briefly describes the classification of data mining techniques and the procedures. This paper also presents the advantage of feature extraction and how classifiers have appropriately been employed. Data mining is useful in extracting knowledge from a huge data set. The paper also suggests future developments for bigger data sets using data mining and also to improve the performance of prediction. In addition, it also suggests water tolerant seed variety and nutrient contents of the soil which can provide better yield to the crop. The paper brings the drawback that there is no unified approach used in data mining where all factors can be utilized for predicting the crop.

E. Manjula, S.Djodiltachoumy (2017) also suggests using data mining techniques so that the data is analyzed from different dimensions and angles. Therefore the data can be converted into historical patterns and future trends. The paper proposes data mining techniques to predict the crop yield production based on the association rules. The paper also briefly describes the analyzing of crop yield based on available data. The proposed work is tested and collected agriculture data obtained for the years from 2000 and 2012. The data in the proposed work is converted into binary values and is mined into frequent pattern in each cluster. The accuracy and error prediction result is also proven and aims at a higher rate.

Monali Paul, Santhosh K Vishwakarma (2015) describes the behavior of soil and predict the yield of a crop using Data Mining Approach, which helps the farmers to select the crops for sowing. In this approach we also use Naive Bayes and Nearest Neighbour methods. These two methods are applied to the soil dataset which is taken from soil testing lab of Jabalpur in Madhya Pradesh. Its accuracy is obtained by evaluating datasets. Both the algorithms are run separately over the training dataset and their performance in terms of accuracy evaluated along with prediction done in testing the dataset.These experiments are performed using Rapid Miner, where it is a software developed by the company of the same name that provides an integrated environment for Machine Learning, text mining, data mining, etc. Rapid Miners use a modular concept, where respective operators have input and output ports through which the operators can communicate with other operators to receive input data or pass the data and generate models over to following operators. This is how the entire analysis process creates a data flow. The experiments are performed from the real world data set obtained from the lab. And the classification of soil is defined into three categories that is low -L , medium-M , high- H.These categories of soil helps in predicting the quality of soil according to the values of nutrients and micronutrients present in it.And finally in the result we observe that categories having maximum confidence value is predicted as the category of that particular soil. Finally from this paper we conclude that the classification of soil into low, medium, and high categories are done by adopting data mining techniques to predict the crop yield using available dataset. This helps the farmers to decide sowing in which land gives the best result for crop production. The future study of the paper aims to create more efficient models using other data mining techniques such as support vector machine, principal component analysis, etc. The major drawback from this study paper we came to know that it uses small data set to the occurrence of some complexity .Hence in the later work there are more chances to use larger dataset of IGB.

A Suresh, P. Ganesh Kumar, etc (2018) proposes a prediction method for the major crops of Tamil Nadu using K-means and Modified K Nearest Neighbour (KNN).The number results shows that our method is better than traditional data mining approach. The main concept of this project is to predict price of major crops of Tamil Nadu region. Though we used three algorithms here K-means, modified KNN, and Fuzzy .But KNN and Modified KNN has proved to be best of all three, as a classification method for enhancing the performance of K- Nearest Neighbour is proposed which uses robust neighbours in training data. Here data mining technique is also used to help farmers in decision making. The main Idea of the presented method is assigning the class label of the data according to K validated data points of the train set where the validity of all data samples in the train set are computed first, and then weighted KNN is performed on any test samples. In this the major drawback we have is they either mainly aim on one crop or predict one parameter like either yield or price. The future study of this paper depends on various Bio inspired methods and provide a comparative study based on the accuracy of each algorithm.

III. SYSTEM METHODOLOGY

The Research methodology mainly has two phases Training and Testing Phase.

IV. TRAINING PHASES

The Training phase was carried out in six phases: Data collection, Data processing, Data visualization, Build the model, Training the model and Test the model.

A. Data Collection

Data collection is the process where information gathered and measured on variables of interest, it establishes a systematic fashion where one answers to a stated research questions, test the hypothesis and evaluates the outcomes. The data sets are collected to build an effective predictive model. This data set contains data of production rate,state,and profit, GDP, Outcome and District.

B. Data Processing

Data processing are a series of actions which are performed on data to verify, organize transforms, integrate and extract data in an appropriate output form. Methods of processing must be rigorously documented to ensure the integrity and utility of the data. The data set that we have collected contains text format in few columns, so we first have to process this text to numerical format. We process the data because we need all the columns of the data set to contain a similar value so we can evaluate the dataset more efficiently. We also make use of null analysis to check if the data set consists of any null values or not. If the dataset contains any null value it can be removed using the null.

C. Data Visualization

Data visualization is an easy way to represent more complex data in the form of graphics. We plot the graphs based on the dataset present to get a clear idea about which group is getting affected by autism. This helps to analyze the data collected. It is used to show the relationship among datasets. In our project we make use of three graphs plotting based on

1. Crop State and region crop
2. Output Production Growth
3. Output Income

Data is split into prime and text format. The first graph plotting that is the output versus State and Crop Growth due to that year. The second graph plotting that is the output versus Production Growth to State region. The third plotting output versus genetics tells us whether group of state region to income stage.

D. Build the Model

After collecting all the necessary details about the model we are interested in designing, we start the process of building the model. Building the model has a few stages in which it is carried out. By building the model it makes it easier to communicate about it with the people and make them understand about the working of our predictive model. Our model is designed to predict if the user enter the crop and district and we have to know about the details of climate, weather and production growth. To develop the prediction of crop yield, the algorithms where built and their accuracy was tested**.**

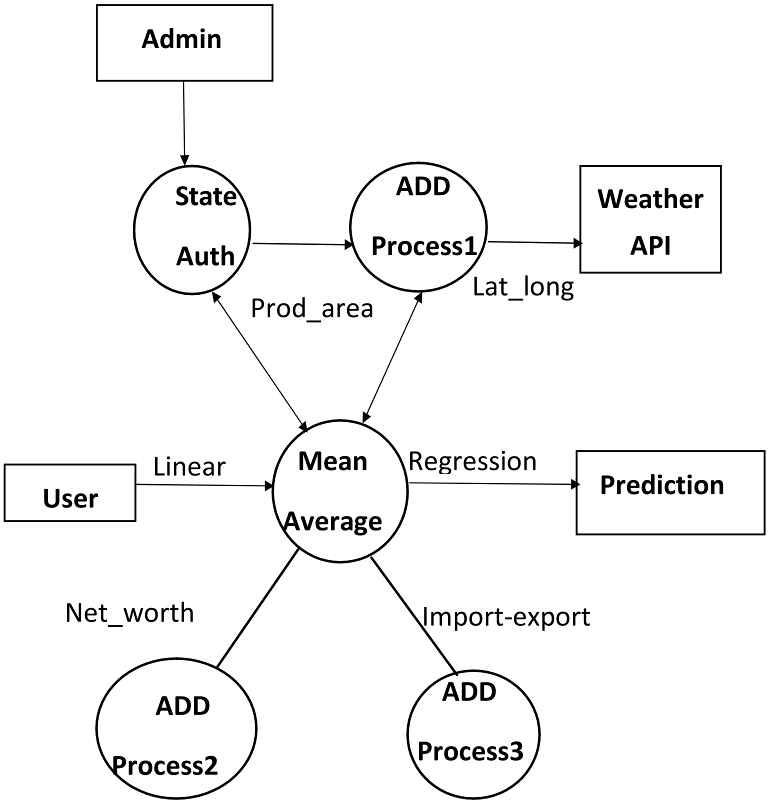

Fig.1 Data flow diagram of the prediction system

E. The Model

Training the model is an important phase in ML. the result we obtain from the model depends on how well we train our model. The performance increases with more than 1000 records. So our model is well trained with all the possible cases. As we have more no of crop yield record the model is trained well with all the data possible. We make use of 70% of the dataset to train the model

F. Algorithm Theory

Linear Regression is a machine learning algorithm based on supervised learning**.** It performs a regression task**.** Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.It is also used to predict dependent variable based on one or more independent variable.

Simple Linear Regression Formula:

Hypothesis Function For Linear Regression

$y = \theta_1 + \theta_2 . x$ Or y=mx + c

Multiple Linear Regression

Y=m1x1+m2x2+….+c

While training the model we are given:

X: input training data (univariate – one input variable (parameter), Y: labels to data (supervised learning).

When training the model – it fits the best line to predict the value of y for a given value of x.The model gets the best regression fit line by finding the best θ_1 and $\theta_2$values.

$\boldsymbol{\theta_1}$**:**intercept

$\boldsymbol{\theta_2}$**:** coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

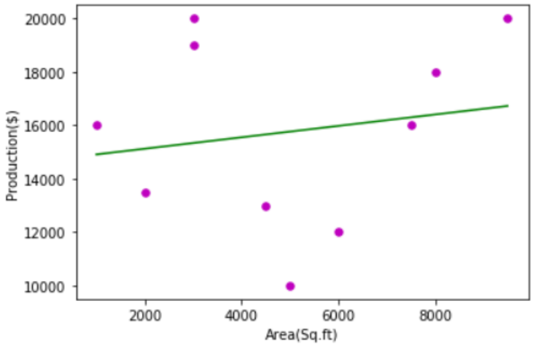

Fig.2 Analysis based on area versus price

From the graphs it is very clear that Linear Regression algorithm is based on independent variables which predict the dependent variables and predict the fit line. Based on the input feature we can predict the output.

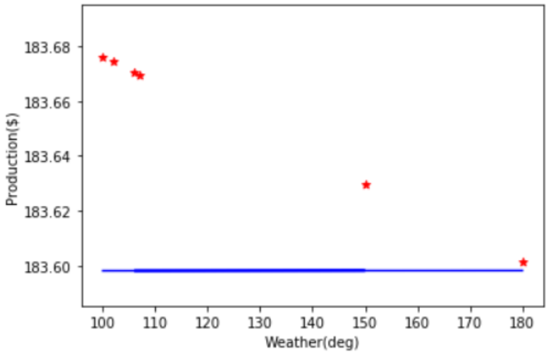

Fig.3 Analysis based on weather versus production

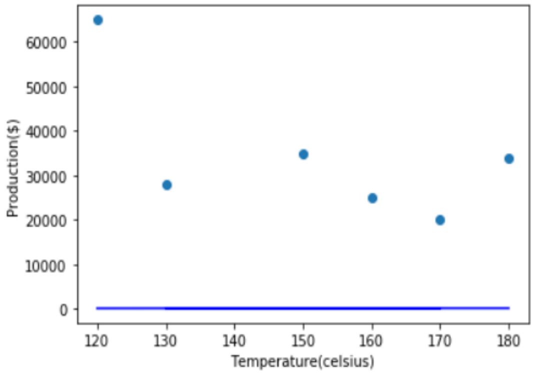

Fig.4 Analysis based on temperature versus production

G. Testing the Model

After training the model with the data set, we then can test the model. We select few set of data and feed the input to the model and check if the model is working well. As we use 70% of the dataset to train we made use of the rest 30% to test the model. Where we can easily get to know if our built model is trained well or not as we already have the prediction to check the output obtained.

V. TESTING PHASES

It consists of loading the trained example model, Get new Crop Yield Prediction record, feeding the record in trained model and Display the result. A Crop Yield Prediction record is been fed and stored and training the data takes place. After training the model with the data set, we then can test the model. We select few set of data and feed the input to the model and check if the model is working well. As we use 70% of the dataset to train we made use of the rest 30% to test the model. Where we can easily get to know if our built model is trained well or not as we already have the prediction to check the output obtained.

VI. RESULT AND ANALYSIS OF CROP YIELD PREDICTION

By using Linear Regression algorithm which is a part of machine learning the crop yield prediction can be done in the prediction system. For that we take crop prediction dataset in that we have n number of input features like crop, production rate, GDP, NDP previous year production growth value. Linear Regression is a Regression technique we are going to predict continuous variable output. Using this algorithm we collect the data and analyze the data. Later the data wrangling is done which is used to clean the data and after that feature selection is done which is used to split which is input features and output. And after that the algorithm is initialized to predict which is the best climate, humidity range to cultivate crops. Because each and every year ranges values for season, weather are different that’s why we are using regression technique.

VII. CONCLUSION

The project " Linear Regression Approach to Predict Crop Yield” is a very good effort to solve the problem of cost loss by creating a machine learning prediction model. In this two modules are very important that state authority and end user module. Here data place a vital role because the above prediction model gives the output based on the input data. Here data consist the information of crop production of past years. And it include following information that is production area of the crop, weather information, So you will get more accurate results when your data is good so that you need to filter and remove all the unwanted data and make a dataset file in csv format from which necessary data will be consider for the computation. Collecting the data for giving the input to project was the biggest challenge because agricultural data is confidential so it is not easy to collect it. This is a generic project because still it can be enhanced and those things will be seen in future enhancement. By looking at all these things I want to conclude that during the development of this project I got an opportunity to learn some important things like how to work under pressure and how to complete the tasks within a given time and so on which help to increase my skills and make it more stronger.

REFERENCES

[1] S.Nagini, DR.Rajinikanth and V.Kiranmayee, *Agriculture yeild prediction using predictive analytic techniques*, 2016.

[2] S.Veenadhari, DR.Bharat Mishra and DR.CD Singh *Soybean productivity modeling using decision tree algorithm*, 2011.

[3] Ms Shreya V Bhosale *etc.al Crop yield prediction using data analytics and hybrid approach*, 2018.

[4] S. Bhanumathi *etc.al*, *Crop yield prediction and efficient use of fertlizers*, 2019.

[5] Yogesh Gandge and Sandhya, *A study on various data mining techniques for crop yield prediction*, 2017.

[6] E. Manjula and S.Djodiltachoumy, *A study on a model of prediction of crop yield*, 2017.

[7] MonaliPaul and Santhosh K Vishwakarma. *Analysis of soil behaviour and prediction of crop yield using DataMining Application*, 2015.

[8] A. Suresh and P. Ganesh Kumar, *etc*,*Prediction of major crop yields of Tamil Nadu using K-means and Modified KNN*,2018