PREDICTION OF CROP YIELD USING GRADIENT BOOSTING

Puneeth Puligudla#1, Kosuri Sai Karthik*2, K V Narendra Kumar#3, Mythili Thirugnanam#4

[1]*School of Computer Science and Engineering, Vellore Institute Of Technology, Vellore, India*

[2]*School of Computer Science and Engineering, Vellore Institute Of Technology, Vellore, India*

[3]*School of Computer Science and Engineering, Vellore Institute Of Technology, Vellore, India*

[4] *Associate Professor (Senior), School of Computer Science and Engineering, Vellore Institute Of Technology, Vellore, India*

Abstract— **India is very well known as an agricultural country because more than 79% of Indians are purely depended on agriculture, and selected it as their occupation. So, it is undoubtedly known that most of the economic growth of the India is relied on products that have been evolving from the agricultural sector. Data mining has acquired a great pickup with the disclosure of computer science and AI, freshly it has been used to project the yield of the crops that has been harvested in the agro-sector. It has been proven that DM algorithms have succeeded in this research. The main motivation to perform this algorithm is that most of the work in the field is manually done so that the people who work in the farm has to suffer a lot. Hence gradient boosting and regression algorithms is proposed so that the harvesting methods are recognized quickly by an automated system which decreases the strive of the farmers. The forecasting is mainly based on climatic changes, the estimation of yield of the crops, pesticides that may destroy the crops growth, nature of the soil and so on. The set of data of these attributes can be predicted using the regression technique. This technique plays a major role in detecting the crop yield data. Results are analyzed and conclusions are drawn as mentioned later in the paper.**

***Keywords*—- AI, Data mining, Crop yield, Regression, Gradient Boosting.**

I. INTRODUCTION

The 3rd largest country In Asia according to present records is India. The development of India is majorly depended on the agriculture. But this has become little difficult now, because most of the conditions that are useful for the growth of the crops in the agricultural field are climate, soil and fertilizers/ pesticides which are changing their nature frequently. Now-a-days climatic changes has become often common and no one can predict what is going to happen next. In agriculture different crops give their yield in different seasons. In this era the floods and waterlessness has increased dramatically and causing lots of loss for the agriculture sector, due to which the farmers are facing a lot of difficulty. Wheat will be grown in the clay loam soil, maize grows in loamy to clay loam soil and the cotton grows in black soil, in this manner different crops grows in different type of soils and in various seasons. But when the scarcity comes into the frame all the efforts of the farmers will be useless. In order to overcome this kind of acts researches has suffered a lot to know about the conditions and preconditions to the crop yield. One of the most important aspect for the crop yield is fertilizers/pesticides, the quantity of using these fertilizers matters a lot in the yield of the crop. Therefore, the use of fertilizers should be adequate this may be bit difficult for the farmers, if mistakenly the quantity is disturbed. Simply we can say that agriculture is a trade that has lot of danger. The danger is in the form of climate, financial, geological... etc. some of that risks can be solved statistically, mathematically or by using some latest computer-based technologies. The development in the agricultural field is only possible through data mining techniques. The databases are very large. Now-a-days the DB can store some millions of data, the size of the data was measured peta and exa bytes in addition to giga and tera bytes. The value for the databases is growing with time. So, the decision should be made very logically. Data mining is nothing but the extraction of knowledge, by using the data that are in large quantity we convert them into usable patterns. The regression technique makes use of the dataset that is stored in the database and form them into a useful equation. If we want to calculate multiple objects then we can go with multiple regression and can create more complex models. There are many types in data mining but spatial data mining among them is used to bring out compulsive patterns in the field of agriculture. Some of the methodologies of data mining can clearly stick to a particular detection in agricultural farm, the k-means methodology can test the soil and classify it based on the navigation(GPS) and is also used for yield prediction, k-nearest neighbor helps to precipitate the weather every day, neural network is used for rain forecasting and speculate the climate, Fuzzy set helps to ferret weeds, the Bayesian network has been developed specially for the agricultural purpose itself. Research says that by using all these methodologies the land quality, rain and so on the crop that can be grown in a specific area can be discovered easily. So, that the crop can be tested in different frames in future and can help the farming. In this way the necessity of a particular plant or crop can be known by making use of the datamining. Hence the farmers hard work will not be wasted and also the development in the agriculture sector increases. The main aim of datamining in agriculture is efficiency and feasibility.

II. Literature Review

A set of research articles have been classified upon the different methods for differentiating the different yield predication systems. E. Manjula and S. Djodiltachoumy[1] proposed that in order to make produced crop much quality. yield checking is the best way they have used data mining techniques and evaluated the future years crop prediction they mainly focused on prediction model based on association rules using these techniques they have uncovered patterns using data which might help farmers to identify crop losses.[1]L. Robert, E. P. Dadios, and Bandala, Argel's [2] through their examination built up the imaginative arrangement that gave a feasible way for detection of diseases in tomato plants. An image catching box that used engine controls was made to capture each tomato plant in all four sides to distinguish and perceive diseases acquired by leaves.[2]

The present Machine Learning paper focused on predicting the crop yield of a specific type on the by using the Random Forest algorithm that provided farmers get a taught on the yield they can expect even before going onto the field proposed by P. Priya*, U. Muthaiah & M. Balamurugan.[3] S. Srinivasan, P. Malliga[4] states that Generally, the ANFIS is very effective for problems having more than 5 or 6 variables. Their work focusses on area selection followed by identification of yield attributes that for the various parameters for prediction. Later the ANFIS model was developed followed by training, testing and validation of data until results are obtained.[4]

Modified DBSCAN method was used to gain knowledge about the cluster data that focused on the areas which had similarities in environmental conditions, soil and rainfall. The authors had used the Partition around Medoids (PAM) algorithm and the Cloud Learning Autonomous Reacting Algorithm (CLARA) techniques to cluster the data based on which the maximum crop produced was predicted by J. Majumdar*, S. Naraseeyappa, S. Ankalaki.[5]. D. Ramesh and V. B. Vishnu[6] in their work made conclusions using multiple linear regression for developing models using certain predictors or factors. The similar values are grouped into a cluster using density-based clustering technique.[6]The system that was proposed by G. Ravichandran and R. Koteeshwari[7] in their paper used the Artificial Neural Networks. They drew their conclusions and predictions were made taking into consideration a few parameters to achieve desired outputs. These parameters included pH, phosphate, potassium, nitrogen, depth, temperature and rainfall.

Plant leaf disease differentiating by machine vision and with the fuzzy logic so that this system efficiently included Information and Communication Technology (ICT) in a harvest and hence it contributes to Precision Agriculture.[9]

.

Various methodologies for detection of edges (edge detection) were worked on, and then they were tested and a comparison was made among them so as to get the best suitable method that can be used for detectiong edges (edge detection). They were namely 1.Roberts Method 2. LoG Method 3. Zerocross method 4. Canny method 5. Sobel Method 6. Prewitt Method had been demonstrated by N. B. A. Mustafa in his paper on image processing on agriculture produce.

III. Methodology

1) Linear Regression

For predicting the crop yield, we will take the choice of crop a farmer would like to grow, the geography of the location, the temperature, rainfall, soil texture and condition and previous data into consideration and try and get the best output possible.

For this we can use the linear regression model and various other machine learning techniques for prediction the best output we can expect.

The fundamental aim of a linear regression model is to get a linear relation between a set of parameters we give as input and the value that we are desired to be predict.

The function of a Linear Regression model is

$y = (\beta + \alpha) * x$ -----------------------------------(1)

In Equation (1), ‘x’ is an input feature, ‘α’ & ‘β’ denote the parameters that we have used to draw inferences regarding the relationships in between x & y. With Linear regression multiple modules can be generated for the same set of data. A large number of various values of ‘α’ & ‘β’ can be given to them and many different lines can be generated to find the relationship in between the prediction value y and the input parameters x.

For measuring accuracy of the model, a Cost Function is used that specifies the error of our model. The difference between the output prediction y’ and actual value of y for each data sample is known as the Error of the Model.

Thus, the Cost Function Can be written as

$J = 1/2m \sum_{i=1}^{m} \left(y'^{i} - y^{i}\right)^2$ -------------------------------(2)

In Equation (2), m is the number of times, various values of ‘α’ & ‘β’ are assigned, y’ is the output prediction and y is the actual value.

However, this takes a lot of time hence we used the Gradient Descent to reduce the Cost Function. We aim at finding a model with lowest cost rate. Manually checking takes a lot of time hence, we used the Gradient Descent Function that works by incrementing ‘α’ & ‘β’ till, a convergence was reached and further incremental change for α & β could not minimize the error severity of the model. In the Linear Regression Model, values of α & β that were figured out or obtained had to be the most optimal values.

$$\alpha = \alpha - \mu * \frac{1}{m} \sum_{i=1}^{m} (y'^{i} - y^{i}) \text{ , } \beta = \beta - \mu * \frac{1}{m} \sum_{i=1}^{m} (y'^{i} - y^{i}) * x^{i}$$

By performing a correlation analysis on the data at-last it was incurred that a strong linear correlation could not be found among most of the parameters and the desired target variable. It was also noted that a few parameter were correlated linearly with the target variable.Through the above made observations, it was concluded that purely linear regression algorithms would not be the most suitable models for prediction. Hence, we decided to use the gradient boosting(Gradient Boosting Trees (GBTs)) algorithm – one of the favorites among many machine-learning algorithms.

2)*Gradient Boosting Algorithm*

One’s whose performance is slightly better than random chance is known as a weak hypothesis or weak learner. Boosting is defined as a process of transforming the mentioned weak learners into strong learners using decision trees fundamentally. A decision tree is trained by assigning equal weights to observations. The weights of the first tree are increased after evaluation if the observations are difficult to classify else are lowered.

This weighted data is used to grow the second tree in an attempt to increase and improve predictions of the first tree. These trees are merged and made as a single prediction tree. From the resultant big tree, we develop a model and grow tree repeatedly for a specific number of iterations to predict the new residuals. The lately developed trees helped in classifying data that were left unclassified by the previous iterations of trees. The weighted sum of the predictions made by the previous models is the final model prediction.

Models using the GBT's are trained additively, gradually and sequentially. The shortcomings of the weak learners are identified using the decision trees. Gradients are used in the loss function ($y= mx+n+e$, e - error). The proposed loss function, can be used to measure the fitting of quality of coefficients of the model on the underlying data.

Using the Gradient Boosting, optimization of a specific cost function, can be allowed rather than that of a loss function that generally does not relate with real life application problems. GBT performs well while working on unbalanced data like real time data management.
This algorithm has a decent number of parameters that are obtained by performing a distribution before-hand known as hyper-parameters. A little tuning of these hyper-parameters can increase the performance exceeding that of the random forest model. Overfitting was one of the most challenging tasks faced and it was overcome by increasing the size of the dataset. Yet improved performance can be achieved using larger datasets. These can be overcome by using ensemble techniques.

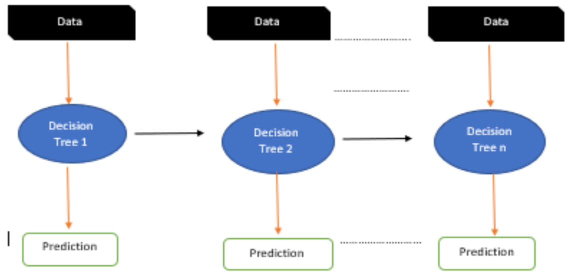

Fig 1. GBT Process

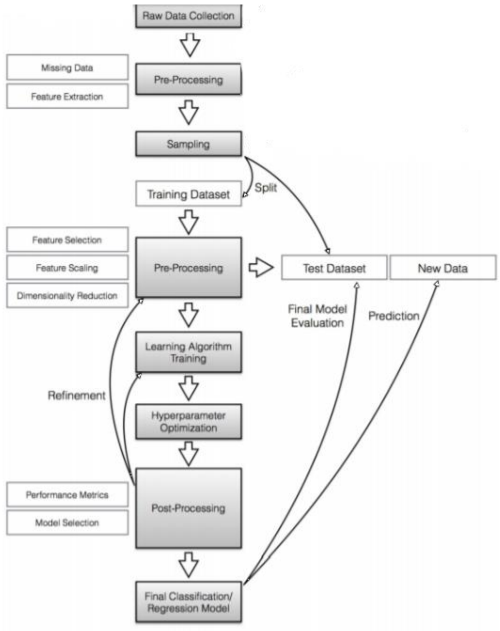

Fig 2. Flow Chart of System

IV. Results and Discussions

Two years of agricultural produce data was taken into consideration and the combinations of hyper-parameters were identified which majorly focused on maximizing the performance. Learning curves and the validation curves were as follows and gave an idea of the bias-variance tradeoff.

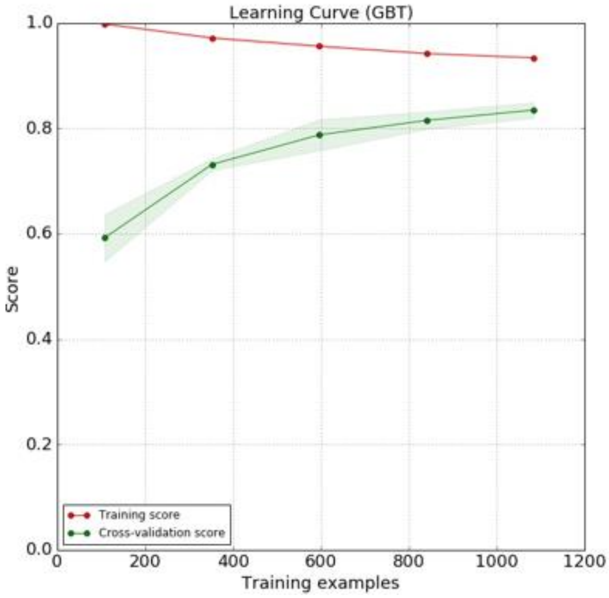

Fig 3. Learning Curve Of GBT

From fig 3 we can incur that the tuned model suggests that there were concerns with regard to overfitting while the variance obtained was decent enough. The overfitting can be tackled using more trained data and a test set was used for the model's performance prediction.

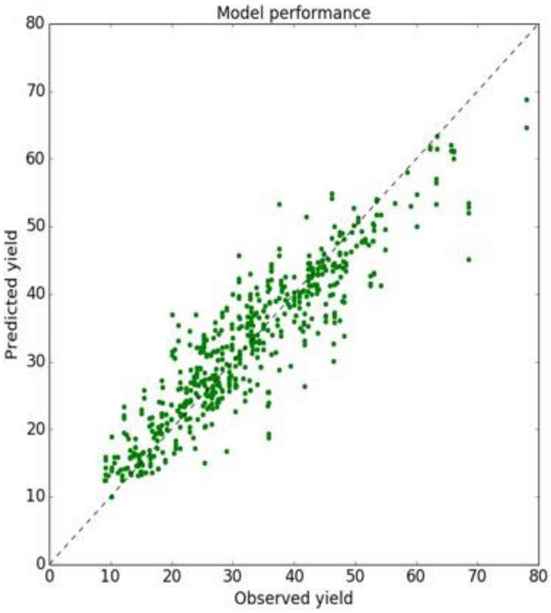

Fig 4 Model Performance

The final model had obtained results with an R^2 value > 0.80 and absolute mean percentage error approximately near to 5%.

V. Conclusion

In this paper, efforts were made to build a system that can make predictions on crop yield. Machine learning’s Linear regression algorithms and Gradient Boost trees were learned and tested on different parameters and datasets during the due course of this research. Simulations were performed for the accurate prediction of crop yield and to assess the system's prediction accuracy.

We predicted the crop yield based on data from multiple sources and using machine learning models like Linear Regression and Gradient Boosting Trees out of which the GBT's proved to be even more efficient. We also found that the prediction accuracy will rely on various factors such as regional difference, type of algorithm used and the agricultural zone. From our work we can conclude that our model is generic and can be used to predict crop yield anywhere in the world.

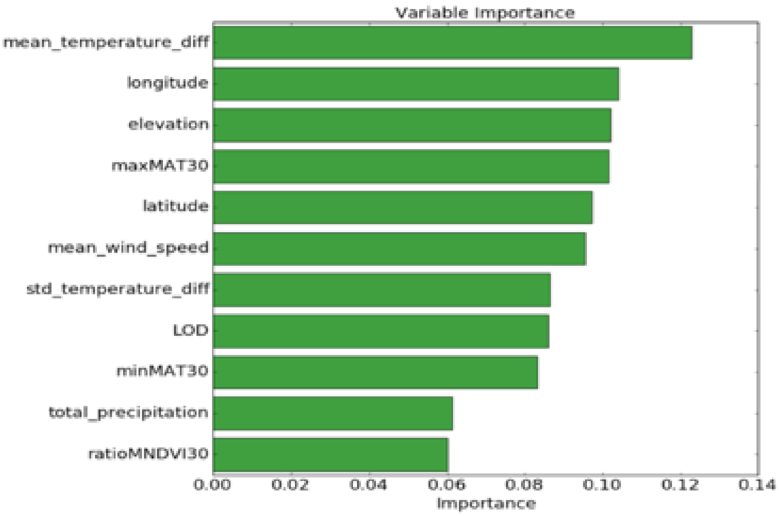

Fig 5. Factors Considered For Prediction and Order Of Importance.

Acknowledgment

Our research wouldn't have been possible had it not been for the guidance of Dr. Mythili Thirugnanam, Associate Professor (Senior), School of Computer Science and Engineering, Vellore Institute Of Technology, Vellore, without whose support and encouragement, this work would not have been a success.

References

[1] E. Manjula and S. Djodiltachoumy, “A model for prediction of crop yield,” International Journal of Computational Intelligence and Informatics, vol. 6, pp. 2349–6363, 2017.

[2] L. Robert, E. P. Dadios, and Bandala, Argel A, “Automated image capturing system for deep learning-based tomato plant leaf disease detection and recognition,” in IEEE, 2018, pp. 1414– 1419.

[3] p. Priya, u. Muthaiah, and m. Balamurugan, “International Journal Of Engineering Sciences & Research Technology Predicting Yield Of The Crop Using Machine Learning Algorithm.”

[4] S. Srinivasan and P. Malliga, “A new approach of adaptive Neuro Fuzzy Inference System (ANFIS) modeling for yield prediction in the supply chain of Jatropha,” in IEEE, 2010, pp. 1249–1253.

[5] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, “Analysis of agriculture data using data mining techniques: application of big data,” Journal of Big data, vol. 4, p. 20, 2017.

[6] D. Ramesh and V. B. Vishnu, “Analysis of crop yield prediction using data mining techniques,” International Journal of research in engineering and technology, vol. 4, pp. 47– 473, 2015.

[7] Sannakki, Sanjeev S, Rajpurohit, Vijay S, V. Nargund, A. Kumar, and Yallur, Prema S, “Leaf disease grading by machine vision and fuzzy logic,” Int J, vol. 2, pp. 1709–1716, 2011.

[8] P. Rothe and R. Kshirsagar, “Cotton leaf disease identification using pattern recognition techniques,” in IEEE, 2015, pp. 1–6.

[9] N. B. A. Mustafa et al., "Image processing of an agriculture produce: Determination of size and ripeness of a banana," 2008 International Symposium on Information Technology, Kuala Lumpur, 2008, pp. 1-7, doi: 10.1109/ITSIM.2008.4631636