A Machine Learning Approach to Crop Yield Prediction

Aditya S Sreerama[1], Dr. B. M. Sagar[2]

[1]*B.E Student,* [2]*Head of Department*
[1,2]*Dept. of Information Science and Engineering, RV College of Engineering®, Bangalore, Karnataka, India*

---***---

Abstract - *Yield prediction benefits the farmers in reducing their losses and to get best prices for their crops. In our current times, owing to unforeseeable climate change, farmers are unable to achieve a reasonable amount of crop production. In order to feed the World's growing population, it is important to integrate new and innovative technologies and resources in the agricultural sector. This Study Focuses on training machine learning models to predict the crop production of the world's most popular crops grown. Factors such as Rainfall, Temperature and Pesticide Input are considered in predicting the crop yield. We compare the accuracy of regression models such as Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor.*

Key Words: Crop Yield Prediction; Regression; Machine Learning

1. INTRODUCTION

Agriculture is one of the most significant factors in the growth of the developing countries such as India where the agricultural ecosystem contributes to about 17-18% of the country's GDP. Agriculture and related industries employ more than 70% of the nation's population and thus is a key source of survival for many. Agriculture also plays a crucial role in the global economy. With the continued expansion of human population awareness of global crop yields is essential to resolving food security issues and reducing the effects of climate change. Crop yield forecasting is an important agricultural problem. Policy makers depend on accurate predictions to pass legislations on import and export policies to strengthen national food security. Farmers also benefit from accurate predictions by making informed strategic management and financial decisions.

Agricultural yield depends primarily on weather conditions such as rain, temperature, etc. and environmental conditions such as Soil Quality, pesticides etc. Accurate knowledge on the history of crop yields is critical for decision-making on agricultural risk management and future predictions.

Although cuisine varies greatly across the world, the essential ingredients that support humans are very similar. The World consumes a lot of maize, wheat, rice and other basic crops. In this study, machine learning approaches are used to forecast the 10 most consumed crops using publicly accessible data from the Food and Agriculture Organization (FAO) and the World Data Bank.

Crop Yield Predicting can be extremely challenging due to the highly varying, non-linear and complex factors that affect it. Added to this, agricultural data is not always collected consistently over large periods of time. It is also very common to find unorganized and incomplete data. In recent times, with increased accessibility to machine learning algorithms, it has become a more reasonable challenge to face. Some of the models that can be used for this kind of prediction include multivariate regression, decision trees, association rule mining and artificial Neural Networks to mention a few.

2. OVERVIEW OF REGRESSION ANALYSIS

Regression Analysis comprise of techniques which leverage a statistical approach to estimate the relationship between dependent variables (also called the 'outcome variable', which is the crop yield in our study) and independent variables (also called 'predictors' or 'covariates', which include weather and environmental conditions such as rain, temperature and pesticide usage in our study) in which the data analyst aims to find a line or other complex linear relationship that fits the given data according to a certain mathematical criterion in a way that does not over fit or under fit the given data. Regression analysis is primarily used for prediction and forecasting in the field of machine learning.

In this study we will compare the accuracy provided by different regression models in predicting crop yield. We measure with a metric called the R2 score. The R2 is a statistical measure which assesses the proportion of the variation in a dependent variable that can be explained by independent variables in a given regression model. The R2 value lies between 0 and 1 where 1 suggests that 100% of the variation in the dependent variable can be explained by the variation in the independent variables.

2.1 Decision Tree Regressor

Decision Tree regressor model is a method commonly used in data mining applications. The aim of the model is to predict the value of a dependent variable based on several independent variables.

The Decision tree iteratively makes decisions on the value of a particular independent variable and continually classifies the dependent variable to make prediction easier. Each internal node of the tree asks a simple question about the value of a certain input feature. Based on the possible outcomes of this question, arcs are drawn to the next decision tree which may ask a question about a different input feature to conclusively predict the value of the feature. Each leaf node of the tree is labeled with a class or a probability distribution over a number of classes. Each node aims to split the data set further to make classification easier and more accurate. This method is called recursive splitting. The splitting stops at a point when further splitting adds no value to the predictions. This learning algorithm falls under the greedy algorithms paradigm.

2.2 Gradient Boosting Regressor

Gradient Boosting is a popularly used Machine Learning Model which uses an ensemble of weak learning algorithms such as decision trees. The model is built step by step with the creation of new decision trees aiming to correct trees built in the previous steps. The trees built in each step are generally shallow. Important considerations of this model include: Number of trees and Depth of Trees.

2.3 Random Forest Regressor

The Random Forest is also called the random decision forest model. This model uses an ensemble of weak decision trees as well. The main difference between Gradient Boosting and Random forests is that each additional decision tree added to the model is trained individually by a random section of the dataset provided. The advantage is that Each of the component decision trees is more robust and the odds of overfitting the model to the data is far lesser that in the case of Gradient Boosters. Important characteristics of this model include: Number of trees and Number of features (independent variables) selected at each node.

3. METHODOLOGY

3.1 Gathering Data

Crop yield data was collected from the Food and Agriculture Organization (fao.org), Crop Yield was provided for more that 200 countries for the world's most common crops. Crop yield was measured in Hectograms per Hectare of land. The crop yield is the target value to be predicted from the machine learning models. The features or independent variables are used in our study include 2 Weather features: Temperature and Rainfall; and one environmental feature: Pesticide usage. The Rainfall data was collected from World Data Bank (data.worldbank.org) which was available which was also categorized country wise and measured in millimeters of rain per year. Average Temperature Data was again collected from World Data Bank (data.worldbank.org) which was measured in Celsius. The pesticide data was collected from the Food and Agriculture Organization (fao.org) which was measured in tons of active ingredient. All these data frames were merged together to for the complete dataset.

3.2 Data Exploration

The final dataset shows data of 101 countries. India is found to have the highest overall crop yield in the world and the highest producer of Cassava and Potato. Potato as a crop is found to be very popular across the world being the highest grown crop in 4 countries. The Yield is considered as the dependent variable while the average temperature, rainfall and pesticide usage are considered as independent variables. To apply suitable learning models, we check the correlation between the columns in the data frame and visualize them with a heatmap. It is evident from the correlation heat map shown in the Figure 1 that all the variables are independent from each other. Figure 2 shows a boxplot of all the crops grown. This helps us understand the distribution of data for each of the crops. We can see from the graph that potatoes, cassava, sweet potato and yams are the world's highest grown crops.

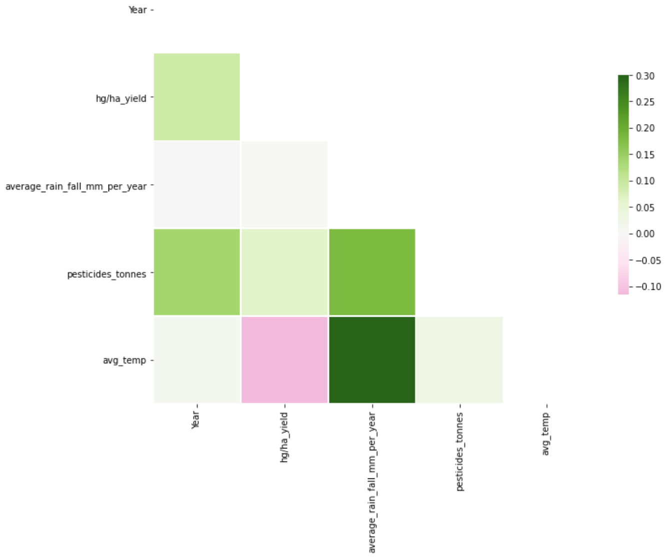


Figure 1. Heatmap showing the correlation between the columns of the data frame

3.3 Cleaning and Preprocessing Data

Typically, data collected from multiple sources are not feasible for analysis this is due to the different data types, units or for- mats that might have been used by different sources. Along with this, datasets may have outliers, wrongly recorded/measured data or missing data. In our current study, outliers did not exist, and rows with missing data were omitted. Object data types were converted to numerical forms to facilitate their use by the machine learning algorithm.

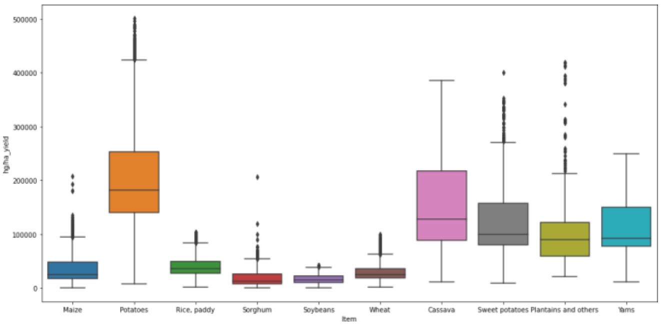
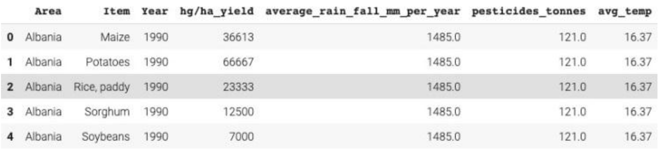


Figure 2. Box Plot to show distribution of data in the dataset



	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
0	Albania	Maize	1990	36613	1485.0	121.0	16.37
1	Albania	Potatoes	1990	66667	1485.0	121.0	16.37
2	Albania	Rice, paddy	1990	23333	1485.0	121.0	16.37
3	Albania	Sorghum	1990	12500	1485.0	121.0	16.37
4	Albania	Soybeans	1990	7000	1485.0	121.0	16.37

Figure 3. A sample of the final dataset used by the machine learning algorithm

3.4 Model Comparison and Selection

In this study, we compare three of the popular regression models: decision tree regressor, Random Forest Regressor and Gradient Boosting Regressor. This is done by training the models with 80% of the dataset and checking accuracy with the rest of the data. The accuracy as mentioned above, is measured with the R2 score. Figure 4 shows the R2 value obtained for the three trained models.

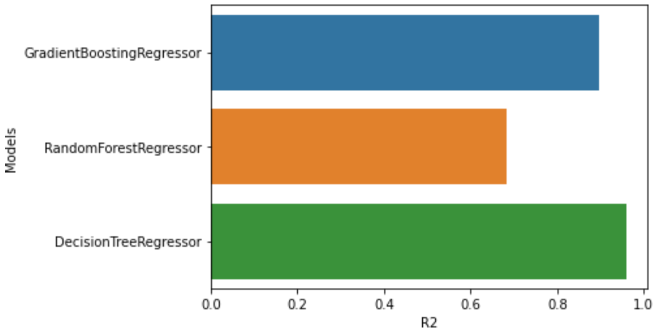


Figure 4. R2 scores of each of the models

3.4 Results

Generally, a higher R2 value indicates a better fit of the model for the given data. It is clear that the Decision Tree Regressor and Gradient Boosting Regressor fit the given data to a very good extent. Figure 5 shows the most important features in predicting the yield. The importance of "Potato" and "Cassava" can be explained by the fact that these two crops are grown in large quantities in many countries around the world. The country "India" forms an important feature because India has the largest crop sum in the dataset as we have explored earlier.

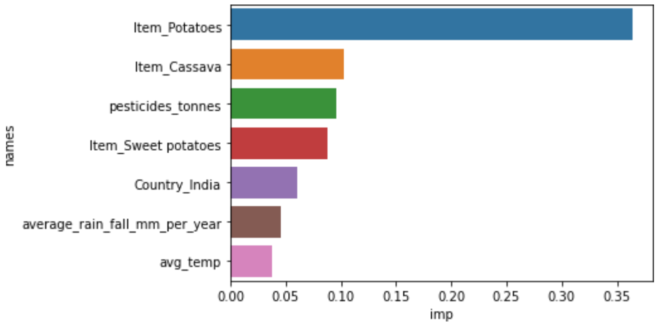


Figure 5. Important features used in prediction

4. CONCLUSION

Crop yield prediction benefits the entire value network associated with agriculture. In this study we use temperature, Rainfall and pesticide usage as the factors in predicting crop yield. The model accuracy can potentially be improved by considering more factors such as climate data, a country's economic conditions, wind and pollution data. We also find that machine learning techniques make it easier to process these large amounts of data and draw actionable information from it.

5. ACKNOWLEDGMENT

This study was supported as a part of the Technical Seminar in the Bachelor of Information Science and Engineering Program in RV College of Engineering®, Bangalore - 560059.

REFERENCES

S. Sunder, "India economic survey 2018: Farmers gain as agriculture mechanisation speeds up, but more r&D needed," The Financial Express, 2018

A. Mucherino, P. Papajorgji, and P. M. Pardalos, "A survey of data mining techniques applied to agriculture," Operational Research, vol. 9, no. 2, pp. 121–140, 2009.

[1] E. Khosla, R. Dharavath, and R. Priya, "Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression," Environment, Development and Sustainability, pp. 1-22, 2019.

T. Horie, M. Yajima, and H. Nakagawa, "Yield forecasting," Agricultural systems, vol. 40, nos. 1-3, pp. 211-236, 1992.

V. Sellam and E. Poovammal, "Prediction of crop yield using regression analysis," Indian Journal of Science and Technology, vol. 9, no. 38, p. 5, 2016.

J. Liu, C. Goering, and L. Tian, “A neural network for setting target corn yields,” Transactions of the ASAE, vol. 44, no. 3, p. 705, 2001.

[2] D. Ramesh and B. V. Vardhan, "Analysis of crop yield prediction using data mining techniques," International Journal of research in engineering and technology, vol. 4, no. 1, pp. 47-473, 2015.