Farmer's Guide: Crop Prediction using Random Forest Regression

SAMRIDH GUPTA[1], RAUNAK JASRASARIA[2], SEEMA KALONIA$_3$, ANSHU KHURANA[4]
[1, 2, 3, 4] *Maharaja Agrasen Institute of Technology*

Abstract- Out of all the three sectors of the Indian Economy, the primary sector has not enjoyed the benefits of technological advancements in recent years as muchas the secondary and tertiary sectors have. Unfortunately, the agricultural sector, on which more than 70% of Indian rural households depend, has been left out of this revolution. Many programs and initiatives have been launched by governing bodies to educate farmers and provide them with technical aid to maximize their harvest. However, not much emphasis has been laid on matching the supply of various crops to their respective market demands. The lack of any such policy has resulted ina surplus crop supply leading to food waste and farmers' money. Our system will guide farmers about how much crop they should produce in a particular year so that there is minimum or no wastage of crops.

Indexed Terms- Random Forest regression, crop production, government data for crop production

I. INTRODUCTION

Agriculture holds immense significance in India as one of the most vital industries. It stands as the most diverse economic sector and plays a pivotal role in the overall development of the entire country. Over sixty percent of the land is utilized foragriculture to feed the nation's 1.3 billion people. Therefore, it is essential to employ cutting-edgeagricultural technology that will benefit our farmers.

Machine learning has been used in agriculture for a while. Crop production forecasting is one of the most difficult issues in precision agriculture, however, numerous models have been developed and are currently performing well. Due to the multitude of factors influencing agricultural productivity, such as climate, weather, soil, fertilizer use, and seed type, this task necessitates the utilization of multiple datasets. Forecasting agricultural yields is a complex undertaking, involving various sophisticated steps. While efforts are continuously made to enhance yield prediction performance, current crop yield prediction models have proven to be reliable in accurately forecasting actual yields.

The most significant of machine learning applications are being developed with the help of a fast-developing approach that aids all sectors in making informed decisions. The majority of modern systems benefit from models being examined beforedeployment.

The primary concept is to increase the throughput ofthe agriculture industry by using machine learning models.

Given that the training period's number of factors was larger than usual, another factor that affects prediction is the degree of information delivered during that time.

This study provides an accurate method for crop forecasting based on historical data collection. Past information on different crop yields and demand is used to provide the information. Our teamdeveloped a program that executes the algorithm and shows a list of crops together with an estimated yield value.

II. LITERATURE REVIEW

Numerous studies have been conducted in the field of agriculture, utilizing machine learning and diverse algorithms to support farmers in crop production.

Girish L [1] discusses the application of machine learning in predicting agricultural output and rainfall. This research delves into several machine-learning techniques for agricultural yield and rainfall prediction. It examines the effectiveness of various algorithms such as linear regression, SVM, KNN, and decision tree. According to the findings, the SVM algorithm emerges as the most accurate in predicting rainfall.

Rahul Katarya [2] outlines the many machine-learning techniques applied to increase agricultural productivity. This study encompasses numerous artificial intelligence techniques, such as precision agriculture through big data analysis and machine learning algorithms. It explores KNN, ensemble-based models, neural networks, and other methods to develop a crop recommender system.

To enhance farmers' profits and elevate the agricultural industry's standards, Ashwani Kumar Kushwaha [3] presents crop yield prediction methods and recommends appropriate crops. This study utilizes the Hadoop platform and an agricultural algorithm to gather extensive data, commonly referred to as big data, for crop production prediction. By analyzing repository data, one can forecast the suitability of a crop for a specific situation and improve overall crop quality.

On datasets from the Indian government, Aruvansh Nigam, Saksham Garg, and Archit Agrawal [4] conducted tests. It was shown that the Random Forest machine learning method provides the best yield forecast accuracy. Simple Recurrent Neural Network, a sequential model, is more effective at predicting rainfall than LSTM is at predicting temperature. For the purpose of yield forecast, the article combines variables such as rainfall, temperature, season, area, etc. When all criteria are integrated, the results show that Random Forest is the best classier.

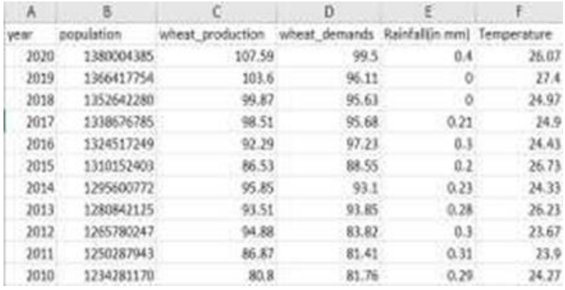
III. METHODOLOGY

In an aim to benefit the farmers, our main motto is to create a machine learning model which will try to predict the market demands of a crop keeping in mind various features such as the previous Year’s demand, production, average temperature, and temperature in the region. We will be using data from the government records like the growing population trend, production of various crops, and their changing demand over the years. This will be used to find the surplus and predict future demand. The data we will take into consideration will be for rice, maize, and wheat as they are primarygrowths for the country. The collected data will be split into training and testing sets, facilitating the model training process and enabling accuracy assessment.

- The data set that we are using is shown below: (Records are from 1960-2020)

The data used is collected from various official government websites and then arranged in a tabular manner, which when converted to a CSV format is used by the Machine learning model.

The data contains the population trend of the country over the years and details like the production and demand of various major crops in India. Using this data and a few other factors the machine learning model calculates the demand for the upcoming years.



A	B	C	D	E	F
year	population	wheat_production	wheat_demands	Rainfall(in mm)	Temperature
2020	1380004385	107.59	99.5	0.4	26.07
2019	1366417754	103.6	96.11	0	27.4
2018	1352642280	99.87	95.63	0	24.97
2017	1338676785	98.51	95.68	0.21	24.9
2016	1324517249	92.29	97.23	0.3	24.43
2015	1310152403	86.53	88.55	0.2	26.73
2014	1295600772	95.85	93.1	0.23	24.33
2013	1280842125	93.51	93.85	0.28	26.23
2012	1265780247	94.88	83.82	0.3	23.67
2011	1250287943	86.87	81.41	0.31	23.9
2010	1234281170	80.8	81.76	0.29	24.27

Fig. 1. Sample of the dataset used for training the model

Population, production, rainfall, temperature are the independent values in the machine learning model whereas the market demand is dependent on these variables.

- Changes in temperature and rainfall over the years due to climate change and its effect on crops

Climate change can exert substantial effects on agricultural production, including alterations in temperature and rainfall patterns that directly influence crop growth, development, and yield.

Climate change can cause changes in the frequency, intensity, and duration of rainfall events. In some areas, there may be more frequent and severe droughts, leading to water scarcity and reduced agricultural productivity. In other areas, more frequent and severe floods may lead to soil erosion and crop damage.

Climate change is causing global temperatures to rise, leading to changes in the timing and length of seasons. This can impact the timing of planting andharvesting, disrupt traditional farming practices, and lead to reduced crop yields. Additionally, higher temperatures can cause heat stress in plants, leading to reduced productivity.

The impact of fluctuating temperatures and rainfall on crop production is contingent upon factors such as the crop type, developmental stage, and duration of exposure to diverse conditions. Some of the specific effects are Reduced yields, Delayed growth, and decreased crop quality.

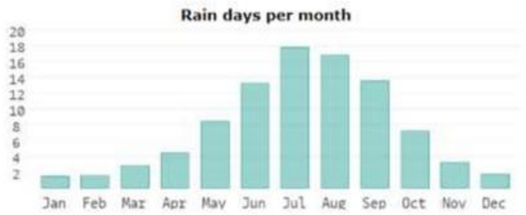


Fig. 2. Avg number of rainy days in different months

- Machine Learning Algorithm for Prediction: Machine learning algorithms play a vital role in prediction. These algorithms generate highly optimized estimations by analyzing input data. Predictive analytics utilizes statistical algorithms, data, and machine learning techniques to calculate the probability of future events based on past data. The objective is to provide the most accurate predictions of future outcomes by going beyond historical observations.

To forecast the production of various crops, we have used a Random Forest regression technique in our system.

Random Forest Regression is a machine learning technique that combines multiple decision trees to create a prediction model for regression problems. Unlike classification, where class labels are predicted, Random Forest Regression predicts continuous numeric values. The algorithm constructs an ensemble of decision trees by randomly selecting subsets of the training data and input features for each tree. This process generates a diverse set of predictors. During the prediction phase, the individual tree predictions are averaged to obtain the final output. While decision trees excel at handling complex data and capturing non-linear relationships, Random Forest Regression takes it a step further by leveraging the collective power of an ensemble of decision trees.

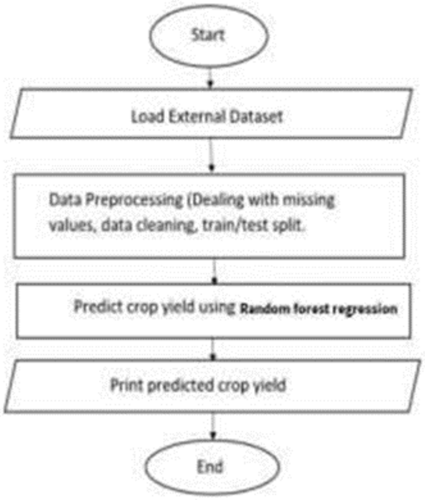

Start
Load External Dataset
Data Preprocessing (Dealing with missing values, data cleaning, train/test split.
Predict crop yield using Random forest regression
Print predicted crop yield
End

Fig. 3. Flowchart for overall prediction of crops

In Random Forest Regression, the algorithm createsan ensemble of decision trees, where each tree predicts a continuous numeric value rather than a class label. The magic lies in the diversity of the trees within the ensemble. To achieve this diversity, the algorithm generates random subsets of the training data and random subsets of the input features for each tree. By doing so, each decision tree learns from a different combination of datapoints and features, resulting in a set of predictors that collectively offer a more robust and accurate prediction.

The process of training a Random Forest Regression model involves several key steps. First, the dataset needs to be prepared, which typically involves preprocessing and cleaning the data, handling missing values, outliers, and transforming variables if necessary. Random Forests can handle a variety of data types, including numerical and categorical features, making them quite versatile.

Next, the algorithm creates random subsets of the original training data, known as bootstrap samples.

These samples are created by randomly selecting instances from the training data, with replacement, to generate diverse subsets. This technique allows for the possibility of duplicateinstances in a bootstrap sample, which contributes tothe diversity of the ensemble. In each bootstrap sample, a decision tree is constructed by recursively dividing the data using different features at each node. At each node of the tree, a random subset of features is selected and used for splitting the data. This process, known as "feature bagging" or "feature subsampling," ensures that each tree focuses on different aspects of the data, preventing overfitting and improving the generalization capability of the model.

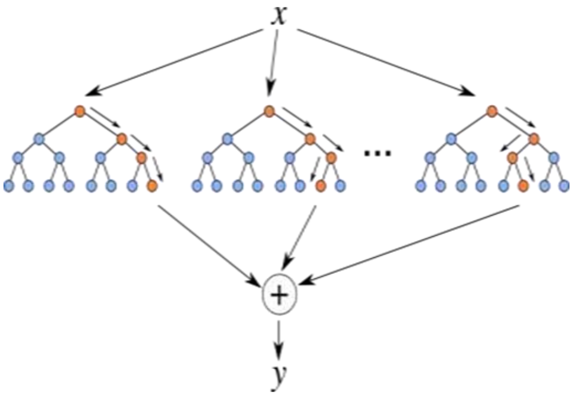


Fig. 4. Random Forest Regression illustration

The training process continues by repeating the creation of bootstrap samples and growing decision trees until a predefined number of trees is reached. The number of trees in the forest is a hyperparameter that needs to be set beforehand. With each treelearning from a different random subset of the data, the ensemble captures a variety of patterns and relationships present in the training set.

After training the Random Forest Regression model, it becomes ready for making predictions. When a new instance requires prediction, it is passed through each decision tree within the forest. Each tree generates a prediction, and in the case of regression tasks, the final prediction is typically obtained by averaging the predictions from individual trees. In some cases, weighted averaging can also be applied to assign more importance to certain trees based on their performance or other criteria.

- The flow chart for crop prediction is shown below:

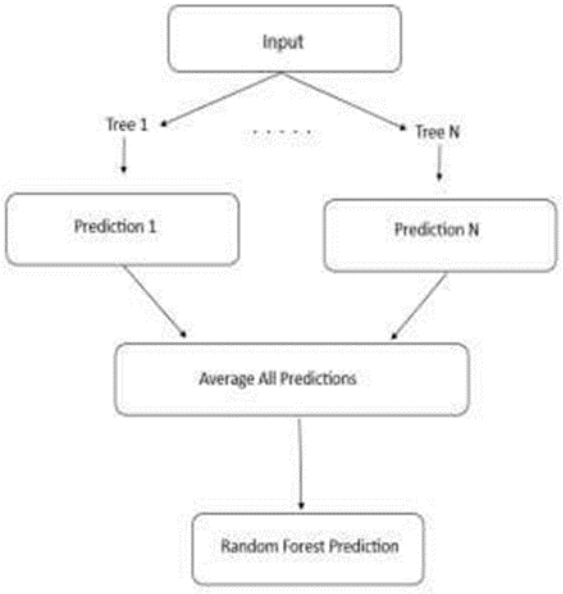


Fig. 5. Flow chart for random forest regression model

IV. RESULT

In order to analyze the performance of a model, there are various metrics available. We have chosen the r2 score metric for accurate prediction. The goodness of fit of a regression model is a measure that evaluates how effectively the model predicts the dependent variable using the independent variable(s). It quantifies the extent to which the model aligns with the observed data and indicates the model's predictive capability.

The R2 score ranges from 0 to 1, where a value of 0 indicates that the model fails to predict the dependent variable, while a score of 1 represents perfect prediction. It serves as a metric to assess the accuracy of the model's predictions, with higher values indicating a stronger fit between the model and the observed data. The model may predictthe dependent variable to some extent if it receives ascore between 0 and 1, with larger scores signifying stronger predictions.

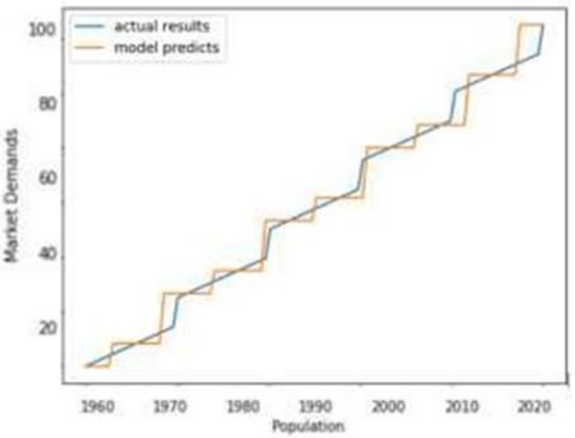
The formula for calculating the R ^ 2 score is:
R^2 = 1 - (SSresidual / SStotal)

In this study, the above-mentioned model a Random Forest was created to find the number of crops required for future consumption. The model was created with an accuracy of 98%.

CONCLUSION

This study focuses on crop forecasting and yield estimation using machine learning methods. Specifically, the random forest machine learning technique was employed to estimate and predict crop production. By analyzing a collection of historical data, a technique was developed to achieve a high accuracy level of approximately 98% in predicting harvests. This approach enables farmers to make informed decisions regarding the selection and quantity of crops to be planted in the field. This work is being done in an effort to understand more about the crops that could be harvested in a useful and efficient way. Accurate predictions of several specific crops in various areas will be advantageous to farmers throughout India. As a result, crop yield rates are increased, which is good for the Indian economy.

At present, farmers are not effectively utilizing technology and analysis, which increases the risk of selecting inappropriate crops and consequently reducing their income. To tackle this challenge, we have created a user-friendly system specifically designed for farmers. This system incorporates a graphical user interface (GUI), allowing farmers to interact with the technology easily. This system enables the prediction of the most suitable crop for a specific plot of land and provides relevant information about necessary nutrients, recommended seeds for cultivation, anticipated yield, and market prices. By offering such comprehensive guidance, our system aims to inspire farmers to make informed decisions for optimal outcomes.



FUTURE SCOPE

Machine learning shows immense potential in the agricultural domain, particularly in areas like crop prediction and improving crop yield. Its applications are diverse and include crop yield prediction, pest detection, and optimization of irrigation practices. These applications leverage the power of machine learning algorithms to enhance agricultural processes and optimize productivity in the field of agriculture. These applications leverage the power of machine learning algorithms to improve agricultural practices and optimize productivity in the agricultural sector.

In the coming years, we can extend this algorithm to include more crops so as to make it more useable and practical in nature. Moreover, we can include various other features which will account for different soil conditions and fertility of different states in the country. We can also make a login pagefor different farmers to log in and make their accounts. We can try implementing data- independent systems as well so that whatever the format, our system should work with the same accuracy.

Overall, the system will be able to cater to the needsof the farmers in the near future

REFERENCES

[1] Girish, L., Gangadhar, S., Bharath, T. R., Balaji, K. S., & Abhishek, K. T. (2018). Crop yield and rainfall prediction in Tumakuru district using machine learning. International Journal for Research in Engineering Application and

Management (IJREAM), 61-65.

[2] Nischitha, K., Vishwakarma, D., Ashwini,

[3] M. N., & Manjuraju, M. R. (2020). Crop prediction using machine learning approaches. InternationalJournal of Engineering Research & Technology (IJERT), 9(08), 23-26.

[4] Kushwaha, A. K., & Bhattachrya, S.(2015). Crop yield prediction using Agro Algorithm in Hadoop. International Journal of Computer Science and Information Technology & Security (IJCSITS), 5(2), 271-274.

[5] Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019, November). Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE.

[6] Mishra, S., Mishra, D., & Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: a review paper. Indian J. Sci. Technol, 9(38), 1-14.

[7] A. Gümüşçü, M. E. Tenekeci, and A.

[8] Bilgili, "Estimation of wheat planting date using machine learning algorithms based on doi: 10.1016/j.suscom.2019.01.010.

[9] Sk Al Zaminur Rahman, Kaushik ChandraMitra, S.M. Mohidul Islam, "Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil Series", International Conference of Computer and Information Technology(ICCIT),21-23 December, 2018.

[10] Yunous Vagh, Jitian Xiao, "Minimum Temperature Profile Data For Shire-Level Crop Yield Prediction", International Conference on Machine Learning and Cybernetics ,Xian, 15-17 july,2012.11.

[11] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. PloS one, 11(6), e0156571.

[12] Zhong, H., Li, X., Lobell, D., Ermon, S., & Brandeau, M. L. (2018). Hierarchical modeling of seed variety yields and decision making for future planting plans. Environment Systems and Decisions, 38, 458-470.

[13] Ahmad, I., Saeed, U., Fahad, M., Ullah, A., Habib ur Rahman, M., Ahmad, A., & Judge, J. (2018). Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. Journal of the Indian Society of Remote Sensing, 46, 1701-1711.

[14] Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016, July). Rice crop yield prediction in India using support vector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-5). IEEE.

[15] https://flask.palletsprojects.com/en/2.2.x/

[16] https://www.data.gov.in