An Overview of Crop Yield Prediction using Machine Learning Approach

Prof. Pritesh A. Patil[1], Mr. Pranav Athavale[2], Mr. Manas Bothara[3], Ms. Siddhi Tambolkar[4], Mr. Aditya More[5]

Dept. of Information Technology
AISSMS's Institute of Information Technology
Pune-1, Maharashtra, India

Abstract - *With the emergence of technologies like machine learning and smart computing, the agriculture field has seen extensive research in recent years. It is getting harder for farmers to use the land effectively to earn the most profit in the unique environment because of the dynamic economics of Agri-produce. Predicting crop yield is a complex task since it depends largely on climatic variables including soil composition, humidity, and rainfall as well as area under cultivation and other necessary metrics. Due to a lack of mapping between environmental variables and accompanying algorithms, many present systems that continuously monitor the aforementioned environmental elements provide inaccurate forecasts. It creates a negative impact on the accuracy of yield prediction. Machine learning techniques are being deployed to precisely forecast the crop output under certain conditions in order to overcome this issue. This project does that by examining and choosing the most accurate machine learning model. In a circumstance like this, when there are several crop alternatives, it is crucial for farmers to prepare their agricultural strategy in advance. The farmer can cultivate in accordance with the crop production estimate if he has it in advance. Crop Yield Prediction (CYP) uses machine learning to assist in making decisions about which crops to grow and how much yield they will produce.*

Key Words: Agriculture, Crop yield, Crop prediction, Random Forest Regressor, Machine learning

1. INTRODUCTION

Machine learning is a rapidly evolving science. Learning is required when we cannot quickly develop a software programme to address a particular problem but instead need example data or experience. Learning is crucial when there is no human knowledge or when individuals are unable to articulate their knowledge. Computers are configured with software to enhance performance standards depending on real or fabricated data. Learning is the application of a computer programme to optimise the model's parameters using training data or past knowledge. We have a model that has been constructed up to a particular point. The model may be descriptive to draw definitive conclusions or predictive to foretell the future [1]. Subfield of AI, Machine Learning provides a way to computers to capture knowledge from the dataset—like chess or generating suggestions on social networks—without needing to be explicitly taught. Agri-tech and precision farming, which are often collectively referred to as "digital agriculture," are developing as new fields of research that employ data-intensive techniques to boost agricultural output while lowering its environmental effect. In agro - based operational contexts, ML has emerged to open up new potential for deciphering, monitoring, and comprehending data-specific processes., along with big data technology and high-end computers. Data analytics sets the groundwork for the development of a diverse array of crop management systems. When data records are involved, perhaps at the scale of large data, fewer ML operations take place. This is primarily due to the additional effort required for the data processing activity, which isn't the case with ML models themselves [2]. Agriculture sector has a major contribution of 15.87% in India's GDP in year 2018-19. Also, it is the principal source of employment in India employing nearly 50% of the population [3]. In addition to being a significant part of the expanding economy, it is crucial for our survival. The primary elements affecting agricultural productivity are climate, infestations, and also capacity of harvesting operations. Having reliable crop history data is crucial for controlling agricultural risk [4]. Immoral and illegal methods are being used to produce higher yields of less-nutritious hybrid cultivars as the population grows. These methods frequently degrade soil quality. It damages the ecosystem. As weather is getting increasingly fluctuant, farmers are unsure about the crop type, correct sowing times and a proper crop strategy. Due to seasonal climatic changes and shifts in the availability of essential resources like soil, water, and air, the usage of various fertilisers is also unclear. In this circumstance, the crop yield rate is steadily declining [5].

Knowing expected yields in advance would help the producer develop a crop strategy. In order to deliver the most practical of its applications, machine learning is a fast-developing approach that aids decision-making across all sectors. Most modern technologies gain from having their prototypes evaluated before being put to use. The major goal is to increase production in the agricultural industry by using ML models. The major emphasis would be on precision agriculture, which puts quality above unfavourable environmental factors [6]. Different performance metrics are evaluated in order to create accurate forecasts and stand by inconsistent trends in rainfall and temperature.

2. LITERATURE REVIEW

A previous study [7] incorporates a data which contains nutrients and other environmental factors, to forecast crops. Various feature selection methods and ML models are used for CYP. The following variables were examined in this study: F1 Score, Mean Absolute Error (MAE), Logarithmic Loss (LL), Accuracy (ACC), Specificity (S), Recall (R), Precision (P), and Recall (R) were used to evaluate the performance of feature selection and classification algorithms (AUC). Six variables—the average soil and air temperatures, min and max air temperatures, precipitation, and humidity—are chosen using the modified elimination of recursive features (MRFE). Different data splitting validation methods like (25-75), (30-70), (35-65), (40-60), (45-55), (50-50), (55-45), (60-40), (65-35), (70-30), (75-25) are incorporated and compared against above mentioned accuracy metrics. Also, variations have been used for feature selection technique in form of MRFE, RFE and Boruta. Results reveal that, among all the aforementioned k-nearest neighbours and bagging classifiers, the Random Forests Classifier provides the greatest accuracy. The values of the measurements fell as the characteristic ranges widened.

Different research [8] forecasts agricultural productivity using a variety of machine learning techniques. The forecasts made by ML models will help farmers choose crop which will produce the highest productions. A technique called data pre-processing is performed to collect the cleaned data from original data collection. As the data is collected in raw form, analysis is not possible. Data is converted into a comprehensible format by using several strategies, such as substituting missing values and null values. The division of training and testing data is the last stage in the data preparation process. Due to the fact that training the model often requires as many datapoints as feasible, the data typically tend to be distributed unevenly. The training dataset, which in this case makes up 80% of the whole collection, is used to teach ML models how to learn and make accurate predictions.

By accounting for factors such as temperature, rainfall, acreage, and other features, the study focuses on the agricultural output of Maharashtra. The most accurate classifier model used in this study is Random Forest, which is followed in accuracy by Logistic Regression and Naive Bayes. Data such as temperature, humidity, rainfall, etc., is fetched using API. The server module receives the data that was retrieved from the API. The server's database is where the data is kept. The user may give information such as location, area, etc. through the mobile application. By completing a single registration, the user may create an account on the mobile app, and all of the submitted information is transferred to the server. The RF model on server-end maps the input to the original data and predicts the output.

Another study by PANDE, SHILPA & RAMESH, PREM & ANMOL, ANMOL & AISHWARYA, B. & ROHILLA, KARUNA & SHAURYA, KUMAR [9] says Farming and allure allied subdivisions are certainly the best providers of livelihoods in country India. The importance of agriculture on a country's GDP cannot be overstated. The enormous extent of the landmass of the nation fortifies it. However, in contrast to universal standards, the agricultural output is unsatisfactory. This is one of the most plausible reasons India's rural farmers have a higher suicide rate. GPS aids in identifying the target area. The input from the customer includes the area and soil composition. ML algorithms compile a list of the most favourable crops, or they forecast the crop yield for a crop that the consumer has chosen. SVM, MLR, ANN, RF, and KNN are some of the chosen ML algorithms that are used to calculate agricultural production. When employed as the ruling class, the Random Forest showed the best results with 95% accuracy. This design makes recommendations on when fertiliser should be applied to improve production. With regard to data sets from specified area, the suggested model forecasts crop yield. The most important factors in predicting present performance are historical data. Several trustworthy sources are used to compile historical data. The data sets are gathered for the regions of Maharashtra and Karnataka. The information includes a number of different parameters, including state, district, year, season, crop kind, area under cultivation, productivity, etc. Other databases with state and district details include the soil type as an attribute. The retrieved soil type column is combined with the primary data set. Similar to how temperature and average rainfall are added to the primary data sets for the particular location, they are acquired from distinct datasets. The data sets have been prepared and cleansed. The mean values are used to replace the null values. Before the algorithms are run, the categorical attributes are transformed into labels. Categorical values in the data sets are dealt with using the one hot encoding method.

The Random Forest regression algorithm was the most accurate of the chosen ones. In order to get the most precise and consistent forecasts, Random Forest constructs numerous decision trees and then combines them. The suggested service for fertiliser usage advises the farmer on when to apply the fertiliser. Using Open Weather API, the model forecasts rain for a specified location for the following 14 days. It advises against using fertilisers if the rainfall is greater than 1.25 mm and is considered "not safe."

A study by Namgiri Suresh, N. V. K. Ramesh, Syed Inthiyaz, P. Poorna Priya, Kurra Nagasowmika, Kota. V. N. Harish Kumar, Mashkoor Shaik and B. N. K. Reddy [10] says the majority of India's agricultural products have been severely impacted by the effects of global warming. Considering their output throughout the previous 20 years. Policymakers and farmers will be able to estimate crop yields early in the harvest by using efficient marketing and storage measures. Farmers will be able to make the appropriate decisions thanks to this technology because it allows them to know in advance approximate yield of their crops prior to cultivation. The machine learning algorithm can then be spread when such a method is implemented with an easy-to-use web-based graphic programme. The farmer is permitted access to the results. However, there are a number of protocols or methods for using data analytics to predict crop yields, and with the help of all those algorithms, we can forecast agricultural productivity. The Random Forest Algorithm is used. The primary goal is to map data on soil and climate characteristics in the dataset that contains yield details over the past 12 years. These factors can help with the prediction of the crops by utilising various classifiers on the given dataset. As a result, several variables are reviewed, and those that strongly support accurate crop prediction are evaluated.

Agriculture is the main driver of economic growth in a developing nation. As a nation's population grows, so does its reliance on agriculture, which in turn affects the nation's upcoming economic growth. Crop yield rates must be increased to address the hunger need of entire country. Some biological measures (such as crop variety, hybrid crops, and insecticides concentration) and chemical ones (such as fertiliser, urea, and potash use) are utilised to address this issue. In addition to such methods, a crop sequencing strategy is required to raise the crop's web yield rate during the growing season. For the purpose of illustrating how it aids farmers in increasing production, the Crop Selection Method (CSM) was utilised as an example. Seasonal crops: Crops can be planted at any time during a season, whereas week-through crops: Crops are also grown all year long. The study demonstrated the usefulness of data mining methods for forecasting agricultural yields based on climatic conditions. Website is user-centric, and all additional grains and regions selected for the analysis should have reliability of prediction above 75%, suggesting higher predictive performance.

A study by [11] says that the right crop must be chosen before being sowed in order to enhance crop output. It relies on a number of variables, including the kind of soil and its makeup, climate, regional topography, crop output, market pricing, etc. The framework of crop selection, which depends on a lot of variables, has a place for methods like Decision Trees, K-nearest Neighbors, and Artificial Neural Networks. Crops have been chosen using machine learning based on how catastrophes like famines might influence them. Artificial neural networks have been used successfully by researchers to select crops based on soil and climate.

To satisfy the demands of the soil, maintain its fertility levels, and subsequently increase crop output, a plant nutrient management system based on machine learning techniques has been developed. It has been suggested to use a crop selection technique called CSM that aids in crop selection based on parameters such as yield projection. The considerable labour that Indian farmers must endure, such as crop selection, irrigation, and harvesting, might be lessened with an accurate weather forecast. Due to the digital divide, farmers have limited access to the Internet and must rely on the few information about weather forecasts that is accessible. Data mining is frequently used to address difficulties in agriculture. Large data sets are analysed using data mining to find valuable classifications and patterns. The main objective of the data mining process is to take the information from a data collection and organise it so that it may be used in other ways. Using the data at hand, this research assesses agricultural yield output. To increase crop production, the crop output was predicted using the data mining approach.

In a study [12], the suggested method tries to anticipate or predict crop production by learning from the historical data of the farming field. Using machine learning techniques, the system constructs a forecasting model by taking into account many variables such as soil characteristics, rainfall, temperature, yield, and other things. Here, we employ a variety of machine learning methods, including decision trees, polynomial regression, and random forests. Predicted accuracy is used to evaluate performance.

A study by D. A. Reddy, B. Dadore, and A. Watekar [13] emphasizes on the fact that India is one of the countries that produces the most agricultural goods, yet its farm productivity is still quite low. So that farmers can earn more from the same plot of land with less labour, productivity needs to be raised. It provides solutions such as providing a recommender utilising an ensemble approach with a high proportion of voting methods using random tree, CHAID, K Nearest Neighbor, and Naive Bayes as a classifier to accurately and effectively advise a good crop based on soil data. Factors are taken into account are soil types, soil features, and crop yield data collecting based on these factors are used to advise the farmer on the best crop to produce. Precision agriculture is one such method that can be used at the right time to maximise yields and productivity.

One such method used in these research projects is assembling. among the numerous machine learning approaches being applied in this area. Ensembling, sometimes referred to as committee methods or model combiners, is a data mining technique that combines the strengths of several models to produce predictions and efficiency that are more accurate than any one model could produce on its own. Random forests are a method for classifying algorithmic rules that uses ensemble learning.

This system makes use of the majority voting procedure, which is the most well-known assembly method. Any number of base learners may be employed in the voting procedure. There must be at least two base learners. The learners are picked so that they complement one other and can teach the others. The possibility of a better prediction increases with competition. Using the supplied training data set, the model is trained. Each model independently predicts the class when a new sample needs to be classed. The class that the majority of the students predicted would finally be chosen as the class label for the new sample.

In another study [14] Rainfall, perception, production, and temperature data sets are taken into account when building a random forest, a collection of decision trees that takes into account two-thirds of the records in the datasets. For correct classification, these decision trees are applied to the remaining entries. The test data can be applied to the resulting training sets for accurate crop yield prediction based on the input attributes. The effectiveness of this strategy was examined using the RF algorithm and the dataset. The benefit of the random forest method is that, in contrast to decision tree machine learning algorithms, overfitting is less of a problem with random forests. No trimming of the random forest is required. Machine learning algorithms using Random Forest can be generated concurrently.

Acquired data sets are converted into csv file format in accordance with the procedure, after which those data sets are loaded. Using a split ratio of either 67 or 33 percentage points, or 0.67 or 0.33, the loaded data sets are split into training and test data sets. To categorise the training data and enable the mapping of attribute values to suitable values and list placement. The data sets should then be summarised after determining the Mean and Standard Deviation for the necessary tuple. Calculate the likelihood by comparing the original data sets with the data list that has been summarised. The biggest probability generated is used for prediction based on the outcome. By contrasting the derived class value with the test data set, the accuracy can be estimated.

Another study [15] says the nation's economy benefits from the field of agriculture. However, it lags behind in utilising current machine learning technology. Therefore, all of the latest machine learning technologies and other new methods should be familiar to our farmers. These techniques help to increase agricultural productivity. Agriculture employs a variety of machine learning techniques to boost agricultural yield rates. These methods can aid in resolving agricultural issues. By examining several approaches, we can also determine the yield accuracy. Thus, by comparing the precision of different crops, we can enhance performance. The use of sensor technologies is widespread in agriculture. This study assists in maximising crop output rates. aids in choosing the appropriate crop for selected land and season.

The primary objective of agricultural planning is to maximise crop output rates while utilising a certain amount of available land resources. Numerous machine learning techniques can aid in increasing crop output rates. When there is a loss due to unfavourable conditions, crop selection can be used to minimise the losses. And under favourable circumstances, it can be employed to increase crop yield rates. By maximising yield rates, nations' economies are boosted.

Crop production may be influenced by the region's natural features, such as riverbeds, hilly terrain, or deep places. such as cloud cover, temperature, rainfall, and humidity. It might be peaty, saline, sandy, or clay soil. In soil, you can find copper, potassium, phosphate, nitrogen, manganese, iron, calcium, ph level, carbon, and different harvesting methods. A number of metrics are used for different crops to generate different projections.

3. CONCLUSION

The diversity of features that are mostly reliant on the availability of data were reviewed in the current study effort, and CYP was calculated using ML methods that were distinct from the features. The geological location, size, and crop features were used to choose the features, and these decisions were mostly influenced by the availability of the data collection. However, using more features did not necessarily result in better outcomes. As a result, testing was done to identify the few best-performing characteristics that were also included in the research. Neural networks, random forests, KNN regression approaches, and various ML techniques were also employed for the best prediction in the majority of the existing models. According to the study, CNN, LSTM, and DNN algorithms were employed the most frequently, but CYP still needed development. The current study demonstrates a number of current models that effectively estimate crop yields while taking into account variables like temperature and weather. In the end, the experimental investigation demonstrated how ML can be combined with the agricultural domain to enhance crop prediction combining the consequences of various factors on agriculture, however, feature selection still needed to be improved.

REFERENCES

[1] Ethem Alpaydın, Introduction to Machine Learning, Second Edition

[2] Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine Learning in Agriculture: A Review. Sensors 2018, 18, 2674. https://doi.org/10.3390/s18082674

[3] Sabitha; AJEBA, 19(1): 18-31, 2020; Article no. AJEBA. 62227 A Study on Sectorial Contribution of GDP in India from 2010 to 2019

[4] Jain A. "Analysis of growth and instability in the area, production, yield, and price of rice in India", Journal of Social Change and Development, 2018;2:46-66

[5] Wolfert S, Ge L, Verdouw C, Bogaardt MJ, "Big data in smart farming– a review. Agricultural Systems", 2017 May 1;153:69-80.

[6] Johnson LK, Bloom JD, Dunning RD, Gunter CC, Boyette MD, Creamer NG, "Farmer harvest decisions and vegetable loss in primary production. Agricultural Systems", 2019 Nov 1;176:102672.

[7] S. P. Raja, B. Sawicka, Z. Stamenkovic and G. Mariammal, "Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers," in IEEE Access, vol. 10, pp. 23625-23641, 2022, doi: 10.1109/ACCESS.2022.3154350.

[8] Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Vinu Williams, 2021, Crop Yield Prediction using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCREIS – 2021 (Volume 09 – Issue 13)

[9] S. M. PANDE, P. K. RAMESH, A. ANMOL, B. R. AISHWARYA, K. ROHILLA and K. SHAURYA, "Crop Recommender System Using Machine Learning Approach," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1066-1071, doi: 10.1109/ICCMC51019.2021.9418351.

[10] N. Suresh et al., "Crop Yield Prediction Using Random Forest Algorithm," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 279-282, doi: 10.1109/ICACCS51430.2021.9441871.

[11] E. Manjula and S. Djodiltachoumy, ``A model for prediction of crop yield,''Int. J. Comput. Intell. Inform., vol. 6, no. 4, pp. 298–305, 2017.

[12] Sangeeta, Shruthi G, "Design And Implementation Of Crop Yield Prediction Model In Agriculture",2020

[13] D. A. Reddy, B. Dadore, and A. Watekar, ``Crop recommendation system to maximize crop yield in ramtek region using machine learning,'' Int. J. Sci. Res. Sci. Technol., vol. 6, no. 1, pp. 485–489 Feb. 2019.

[14] Priya, P., Muthaiah, U., Balamurugan, M."Predicting Yield of the Crop Using Machine Learning Algorithm",2015

[15] Ramesh Medar,Vijay S, Shweta, "Crop Yield Prediction using Machine Learning Techniques", 2019