International Journal of
Human Computations & Intelligence
Vol. 03, Issue. 02, Mar/Apr 2024

RESEARCH ARTICLE OPEN ACCESS

Hybrid Mode of Crop Yield PredictionUsing Various Machine Learning Algorithms

Sangeetha Muthu . Callins Christiyana Chelladurai . Hari Nainyar Pillai C

Department of Computer Science and Engineering,
SRM Madurai College for Engineering and Technology, Madurai, India

DOI: **10.5281/zenodo.11440956**
Received: 21 January 2024 / Revised: 06 February 2024 / Accepted: 12 March 2024

Abstract – Over 50% of India's population depends on agriculture for existence, making it the foundation of the country's economy. Variations in weather, climate, and other environmental factors are now a significant threat to the continued success of agriculture. The decision support tool for Crop Yield Prediction (CYP), which includes assisting decisions on which crops to plant and what to do during the growth season of the crops, is where machine learning (ML) plays a vital role. The current study focuses on a systematic review that extracts and synthesizes the CYP traits. In addition, a number of approaches have been created to analyse agricultural production prediction utilizing Artificial Intelligence techniques. In this paper, the predictions provided by the Random Forest and Naive Bayes algorithms will assist the farmers in choosing which crop to cultivate to produce the greatest yield by taking into account variables such as water, wind, sunlight, temperature, rainfall, and photosynthetic activity. Pollinating agents, which analyses several ML strategies used in the field of agricultural yield estimation and offered a complete study in terms of accuracy employing the techniques, boost the fertility of the soil.

Index Terms – Crop prediction, Machine learning, Artificial Intelligence, GDP, Naive Bayes, Random forest.

I. INTRODUCTION

India has a long history of agriculture dating back to the Indus Valley Civilization. In this industry, India is ranked second. 15.4% of the GDP (gross domestic product) is made up of the agricultural and related industries, which employ around 31% of all workers. India leads the world in net cropped area, followed by the US and China. The largest economic sector in terms of population diversity is agriculture, which is important to India's entire socioeconomic structure. Due to the industrialization revolution, agriculture's economic contribution to India's GDP is progressively

MILESTONE RESEARCH.IN OPEN ACCESS

ISSN (Online): 2583-5696
Int. Jr. of Hum Comp. & Int.

falling along with the nation's overall economic expansion. The inability of Indian agriculture to integrate technology to provide the intended results is a challenge. The patterns of rainfall and temperature are thrown off by the introduction of new technology and the excessive usage of non-renewable energy sources. Farmers find it difficult to accurately estimate temperature and rainfall patterns due to the uneven trends brought on by the negative consequences of global warming, which has an impact on their crop production productivity. Different machine learning algorithms, including RNN, LSTM, and others, can be used to obtain a pattern in order to perform reliable prediction and handle inconsistent trends in rainfall and temperature. It will support India's expanding agricultural sector and, taken collectively, improve farmers' quality of life. In the past, numerous scholars have used machine learning approaches to improve the nation's agricultural growth. This Paper combines a variety of machine learning approaches to forecast the crop's output. Based on mean absolute error, these methodologies' results are contrasted. By taking into account variables like temperature, rainfall, area, and other aspects, the predictions provided by machine learning algorithms will assist farmers in choosing which crop to grow to receive the greatest yield.

II. PROPOSED SYSTEM ARCHITECTURE

Although the schooling tool is similar with that of device mastering models, the data-pushed crop version takes a basically and philosophically absolutely one among a type studying method from conventional neural networks. Neural networks use a mean modeling form with a huge quantity of parameters and rely nearly completely on facts to research the enter-output courting without preset underlying assumptions. This method has the capability to capture pretty subtle and insightful statistics past the comprehension n of human intelligence. Along with this capacity advantage come risks. The first is the hazard of data deficiency, every quantitatively or qualitatively, that would mislead the model into collecting biased or fake records and offsetting the functionality gain.

The second drawback is the huge form of parameters, which may be crucial to attain a common approximation functionality, however they make the model no longer simplest liable to overfitting however moreover tough to offer an purpose of. On the alternative hand, the form of the statistics-driven crop model is determined regular with human know-how of plant body structure, that is superior enough to qualitatively describe the crop boom device; historical information were used most effective to calibrate a small variety of biologically meaningful parameters. For example, the reality that radiation contributes to photosynthesis is incorporated in the form of the model, on the equal time as historical data were used to quantitatively determine the precise rate of radiation contribution to photosynthetic yield. These genotypic parameters are independent of environmental influences, for this reason can be used to understand genetic characteristics of particular genotype. The proposed system of this paper is basically followed with the following architecture. In that crop has been collected in various formats and it has been taken into the data storage. Then the data has been inculcated with data from land that can be differentiated and collected in format of text file. This will be loaded into the data storage to make further steps of analysis. This can be applied in two following algorithms are Random forest and Naïve bayes algorithm. The final result will be produced with comparison of two algorithms by considering the number of parameters like Water, Wind, Sunlight, Temperature, Rainfall, Photosynthesis, Microbes.

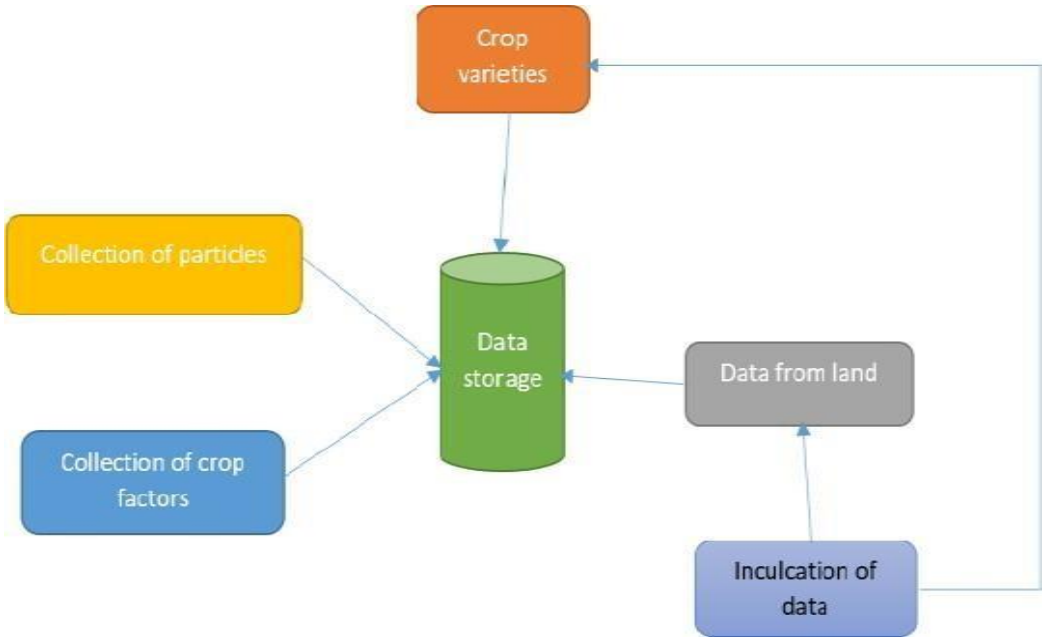

Fig.1 Architecture diagram for crop yield prediction

The Random Forest classification

Random forest is frequently employed and frequently highly effective is random forest. It is a multi-decision tree model ensemble classifier. The various outcomes can be combined using ensemble models. Both classification and regression models can be used with the random forest model. In essence, it splits the data set into smaller sections before processing it. Large datasets can be processed efficiently using random forest models since all computation may be divided, making it simpler to run the model concurrently. Without variable deletion, hundreds of input variables can be processed.

It calculates distances between examples that can be used for clustering, finding outliers, or (by scaling) producing visually appealing representations of the data. The biggest issue with the random forest classifier is its complexity, but there are other drawbacks as well. Working with random forests requires more expertise than using traditional decision trees because it is more difficult. Additionally, the complexity places a heavy burden on computing resources. Organizations in the financial sector frequently employ Random Forest. The forecast of credit risk is a frequent use-case. If you've ever sought for credit, you may be familiar with the inquiries that banks make. Frequently, they are added to random forest models.

The Naive Bayes classifier

Based on past knowledge of circumstances that might be related to an occurrence, the Naive Bayes classifier predicts outcomes. The Bayes Theorem is the basis of it. The presumed qualities are highly independent of one another. It computes ratios between events using categorical data. Naive Bayes

has varied advantages. It can forecast classes of data sets quickly and easily. Additionally, it can forecast several classes. Compared to models like logistic regression, naive Bayes performs better and requires a training data. A significant problem is that the model will give a 0 (zero) probability if a categorical variable contains a category that was not checked in the training data set, which prevents prediction. The term linear version has been created based on multiple linear regression analysis. the non-linear model has been constructed the usage of synthetic neural networks (ANN)to are expecting the potato yield of very early types [7]. Phytophenological and meteorological records have been used for modelling.it used a few forecast blunders metrics which include worldwide relative approximation, root mean square error, imply absolute errors, and mean absolute percentage errors within the validation degree.

Data Pre-processing

A technique for turning filthy data into clean data sets is data preparation. In other words, if data are collected from different sources, they are done so in a way that prevents analysis.

Encoding Categorical Variables

Two category columns in the data frame, which are variables with label values rather than numeric values, are present. In many cases, the range of possible values is restricted to a specified set, like in this case, the values for the goods and countries. A number of machine learning algorithms cannot directly act on label data. They mandate that all variables for input and output be integers. In order to become numerical data, categorical data must be converted. One popular encoding technique entails converting categorical data into a format that may be supplied to ML algorithms to aid in their prediction performance. To do this, the One-Hot Encoding technique will be used to transform these two columns into a one-hot numeric array.

Scaling Features

The aforementioned dataset contains features that span a variety of magnitudes, units, and ranges. In distance calculations, the magnitudes of the features will be significantly more significant than their magnitudes. To lessen this effect, we must balance the sizes of all features. Scaling can support this.

Training and Test Data:

The original dataset will be used to produce the training dataset and test dataset. As many data points as possible are required for the model to train, which frequently leads to data inequality. The standard ratios for train/test are 70/30 or 80/20. The training dataset is the initial set of data needed to train a machine learning algorithm to learn and produce precise predictions. The training dataset makes up 70% of the dataset. The test dataset is nevertheless used to assess how effectively the ML algorithm was trained using the training dataset. It would be useless to evaluate the approach by merely recycling the training dataset because the ML algorithm would already "know" the desired result. 30% of the dataset is a test dataset.

III. COMPARING AND SELECTING MODELS

The potential to proper away and reliably forecast crop yield is an important element of local and global meals protection. Forecasting three hundred and sixty five days-to-12 months versions within the yields of important plants on the nearby and country wide ranges can give a lift to the ability of societies to reply to meals manufacturing shocks and meals charge spikes triggered through excessive activities. Forecast would probably no longer have ensured a special very last effects, it may clearly assist governments improve their making plans and mitigate the potential effects of drought. Within-season crop forecasting is likewise important for farmers to make extra informed crop control and financial picks.

In order to train the version, we furthermore want to have crop yield observations within the path of numerous years and with a bit of pinnacle fortune beneath specific climatic conditions to provide a few examples inside the dataset. With the input variables and the located output we are able to now begin the technique of training the model using ML algorithms. There are several sorts of ML algorithms; generally, multiple mathematical model is produced based sincerely truly totally on the unique algorithms used. To test the models and decide which one is the splendid appearing version, a part of the dataset is left unused and in the long run used to determine how nicely the decided version plays inside the course of the unused information. Once the amazing appearing version is selected, it could be deployed to provide operational crop forecast outputs on a everyday foundation.

Model outputs can be a numeric output or a class which incorporates low, not unusual, or excessive yield. As in any scenario in which mathematical fashions are used to extract data from records, the results can be as first-rate due to the reality the records used to generate them. Crop yield is an essential variable in lots of disciplines. Global yield datasets for the historical beyond have an increasing number of been used to analyze climate-crop relationships, meals manufacturing potential, food deliver and call for, carbon and nitrogen biking, greenhouse fuel emissions from agriculture and land-use trade. Recently, meals manufacturing losses because of weather and climate extremes below converting climate and improved stakeholder preparedness are concerns for many societies as the sector experiences population boom and subsequent increases inside the call for agricultural merchandise. An analysis of climate-crop relationships, specially, the outcomes of climate and climate extremes on food manufacturing, calls for a spatially explicit yield dataset spanning several a long time. At the worldwide scale, this kind of dataset has simplest lately been advanced.

The global dataset of ancient yield for important plants is an example of such a dataset. The GDHY is a hybrid of agricultural census statistics and satellite far off sensing. Crop harvested area maps, crop calendar and percentage of production amount in specific developing seasons for a crop also are used as inputs for the GDHY dataset. Therefore, the grid-mobile yield values recorded in the GDHY dataset are version estimates rather than observations. Since its improvement and initial launch in previous years, efforts have centered on improving the facts high-quality, assessing uncertainties and extending the time coverage to encompass greater recent years. The table is being discussed to compare two algorithms which are all Random forest and naïve Bayes by considering the Following parameters water, wind, sunlight, temperature, rainfall, photosynthesis, microbes. The deviations in

the result

IV. RESULTS

Two category columns in the data frame, which are variables with label values rather than numeric values, are present. In many cases, the range of possible values is restricted to a specified set, like in this case, the values for the goods and countries. A number of machine learning algorithms cannot directly act on label data. They mandate that all variables for input and output be integers. In order to become numerical data, categorical data must be converted. One popular encoding technique entails converting categorical data into a format that may be supplied to ML algorithms to aid in their prediction performance. To do this, the One-Hot Encoding technique will be used to transform these two columns into a one-hot numeric array.

Scaling Features

The aforementioned dataset contains features that span a variety of magnitudes, units, and ranges. In distance calculations, the magnitudes of the features will be significantly more significant than their magnitudes. To lessen this effect, we must balance the sizes of all features. Scaling can support this. explains little variations to get effective results. Finally Naïve bayes achieves higher percentage compared to Random forest in many of the parameters. The original dataset will be used to produce the training dataset and test dataset. As many data points as possible are required for the model to train, which frequently leads to data inequality. The standard ratios for train/test are 70/30 or 80/20. The training dataset is the initial set of data needed to train a machine learning algorithm to learn and produce precise predictions. The training dataset makes up 70% of the dataset. The table is being discussed to compare two algorithms which are all Random forest and naïve bayes by considering the following parameters water, wind , sunlight, temperature, rainfall, photosynthesis, microbes. The deviations in the result explains little variations to get effective results. Finally Naïve bayes achieves higher percentage compared to Random forest in many of the parameters.

The test dataset is nevertheless used to assess how effectively the ML algorithm was trained using the training dataset. It would be useless to evaluate the approach by merely recycling the training dataset because the ML algorithm would already "know" the desired result. 30% of the dataset is a test dataset.

Table: 1: Performance estimation

Parameters	Random forest	Naïve bayes
Water	87.28469172	88.75698
Wind	89.42157938	88.52227
Sunlight	90.463335	91.54978
Temperature	88.79358427	91.55812
Rainfall	90.57931479	91.5135

Photosynthesis	88.99344522	92.79358
Microbes	83.5581168	88.22221

V. CONCLUSION

In conclusion, machine learning-based crop outputprediction has the potential to completely transform the agricultural sector. This technology can assist farmers in achieving greater yields and more successful enterprises by supplying more accurate projections, improved decision-making, increasing efficiency, and promoting sustainability. While there are significant difficulties in applying machine learning to estimate crop yields, the advantages are obvious, and future developments in this area are expected.

REFERENCES

1. Kaur, M., Gulati, H., & Kundra, H. (2014). Data mining in agriculture on crop price prediction: techniques and applications. International Journal of Computer Applications, 99(12), 1-3.
2. Meng, J. (2016). Research on the cost of agricultural products circulation and its control under the new normal economic development. Commercial Times, 23, 145-147.
3. Chen, Y., Lee, W. S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., & He, Y. (2019). Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. Remote Sensing, 11(13), 1584.
4. Kaloxylos, A., Eigenmann, R., Teye, F., Politopoulou, Z., Wolfert, S., Shrank, C., ... & Kormentzas, G. (2012). Farm management systems and the Future Internet era. Computers and electronics in agriculture, 89, 130-144.
5. Costa, L., McBreen, J., Ampatzidis, Y., Guo, J., Gahrooei, M. R., & Babar, M. A. (2022). Using UAV-based hyperspectral imaging and functional regression to assist in predicting grain yield and related traits in wheat under heat-related stress environments for the purpose of stable yielding genotypes. Precision Agriculture, 23(2), 622-642.
6. Ahmed, S. T., Basha, S. M., Arumugam, S. R., & Kodabagi, M. M. (2021). Pattern Recognition: An Introduction. MileStone Research Publications.
7. Chen, Y., Lee, W. S., Gan, H., Peres, N., Fraisse, C., Zhang, Y., & He, Y. (2019). Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. Remote Sensing, 11(13), 1584.
8. Iizumi, T., Shin, Y., Kim, W., Kim, M., & Choi, J. (2018). Global crop yield forecasting using seasonal climate information from a multi-model ensemble. Climate Services, 11, 13-23.
9. Basha, S. M., & Ahmed, S. T. (2023). Real Time Systems: Challenges and Applications.
10. Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific reports, 11(1), 1606.
11. Vijayakumar, V., Costa, L., & Ampatzidis, Y. (2021). Prediction of citrus yield with AI using ground-based fruit detection and UAV imagery. In 2021 ASABE Annual International Virtual Meeting (p. 1). American Society of Agricultural and Biological Engineers.
12. LK, S. S., Rana, M., Ahmed, S. T., & Anitha, K. (2021, November). Real-Time IoT Based Temperature and NPK Monitoring System Sugarcane-Crop Yield for Increasing. In 2021 Innovations in Power and Advanced Computing Technologies (i-PACT) (pp. 1-5). IEEE.