Exploring Precision Agriculture: A Weka-Based Approach for Accurate Yield Prediction in Indian Rice Farming

by

23BDS1153	Sharan K
23BDS1071	Riddhima Singh
23BDS1136	Rounak Handa
23BLC1315	Syed Arsalan
23BEC1075	Kruthika Karra
23BDS1152	Kartikeya Mittal
23BEE1032	Adrija Das
23BLC1248	Suvarna Bhatnagar
23BCE1622	Archit Vishnoi
23BCE1809	Shubham Bhuyan

a project report submitted to

Dr. Ravi Prakash Dwivedi

in partial fulfillment of course

INTRODUCTION TO ENGINEERING

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

VIT CHENNAI

VANDALUR-KELAMBAKKAM ROAD

CHENNAI - 6000127

Table Of Contents

Abstract

"In the past few years, the popularity of machine learning methods has surged alongside the emergence of big data. When utilized as input for machine learning methods, the substantial volume of generated data yields valuable insights. The integration of these methods in the agricultural industry holds significant promise for boosting both crop productivity and quality.

Within this study, we examine agricultural data sourced from a farming site in India, encompassing Kharif cropping patterns. Furthermore, we employ a smart farm ontology housing pertinent concepts and properties within the agricultural domain. We construct a knowledge graph by associating the acquired data with our smart farm ontology. Leveraging this knowledge graph, we derive structural insights and consolidate data using WEKA. Our machine learning models subsequently employ consolidated data to forecast crop yields, offering advantages to farmers and diverse stakeholders. Additionally, we conduct a comparative analysis of the outcomes produced by different machine learning models utilized in our study."

The best overall results were obtained from the Random Tree method on rice dataset with **87.46%**.

Keywords: Machine learning, agriculture, Kharif crops, WEKA, crop yields, knowledge graph, smart farm ontology.

Introduction

Accurate forecasting of crop production holds immense significance in efficient agricultural governance, particularly in economies like India heavily reliant on agriculture. This research centers on employing data mining, an interdisciplinary methodology encompassing artificial intelligence, computer science, machine learning, and statistics, to anticipate rice crop yields in India.

Rice cultivation contributes significantly, constituting over 40% of total crop production. Its success relies heavily on climatic conditions. Enhancing the capability to forecast crop output amidst varying climatic elements like rainfall, temperature, cloud cover, vapor pressure, and potential evapotranspiration can aid farmers and stakeholders in crucial decisions regarding agronomy practices and crop selection. Predicting agricultural yields poses both a challenging and vital task for nations. This estimation often relies on historical crop data and environmental factors, enabling the anticipation of crop productivity, usually measured in kilograms per hectare. Employing diverse data mining techniques facilitates the analysis of such data for accurate yield predictions. Government bodies, policymakers, agro-based industries, traders, and agriculturists extensively benefit from crop yield projections. Governments utilize these forecasts in various aspects such as procurement, distribution, buffer-stocking, import/export, price determination, and agricultural commodity marketing.

Agriculture stands as the cornerstone of the Indian economy, with 70% of the rural population's livelihood dependent on agricultural activities. Crop yields are intricately linked to environmental aspects like precipitation, temperature, and evapotranspiration. Farmers traditionally base their cultivation practices on past experiences, but the growing uncertainty in environmental conditions necessitates a thorough analysis of historical environmental data for

successful farming. Furthermore, analyzing past cultivation records contributes significantly to achieving higher yields. Predicting crop output relies on historical crop and weather data through the application of data mining methodologies.

Related work

Increasing the quality and quantity of the agriculture crop has become a crucial task for the food industry in order to ensure food sustainability for the growing population. There- fore, farmers have started utilizing machine learning (ML) models for managing their farms efficiently by making key decisions with the help of data-driven insights. Earlier re- searchers mostly implemented non-linear iterative multivariate optimization approaches and empirical piecewise linear crop yield prediction equations that make use of vegetation index and meteorological parameters for predicting crop yield .

However, considering the capabilities of ML algorithms such as data-driven insights in real-time scenarios widely used in various CPS domains, researchers have started implement- ing these algorithms in agriculture for better management and efficiency of the agricultural farms. For example, Behmann et al. described how ML models can be used for protecting the crop in advance by detecting the biotic stress . Another example is irrigation recommendations provided by the ML algorithms by utilizing farm data such as soil moisture, weather, etc .

Methodology

In order to explore the possibility of predicting the effect of weather on rice yields in India.Open source Data mining tool WEKA version 3.8.6 was used. Necessary rice yield data and weather data was collected.The yield data of rice and daily weather data were used over a period of 31 years (1988-2019).The averaged data of bright sunshine hours, temperature (maximum and minimum, relative humidity (morning and evening) and total rainfall were used to judge combined effect on rice yields.

Steps followed while predicting rice yields:

1.Initially, rice yields were categorized as L - Low, N - Normal and H – High. Dataset was prepared in Excel sheet with CSV extension for analysis by open source data mining tool WEKA (V3.8.6). Min-Max Normalization technique was used to normalize the experimental dataset which reduces large variation of yield prediction. Feature Selection algorithm viz., 'cfsSubsetEval' was also used as it improves the quality of the data and increases the accuracy of data mining algorithms. It also focuses on eliminating redundant and irrelevant data

2.The raw data is cleaned and sorted.

3.The classifiers viz., MLP and RandomTree were then employed over the trained data.

4. The results of each algorithm were noted from WEKA and compared with each other.

5. Correctly Classified Instances, Incorrectly Classified Instances, Mean Absolute Error, Relative Absolute Error, Root Mean Squared Error and Root Relative Squared Error values were taken into consideration for each case.

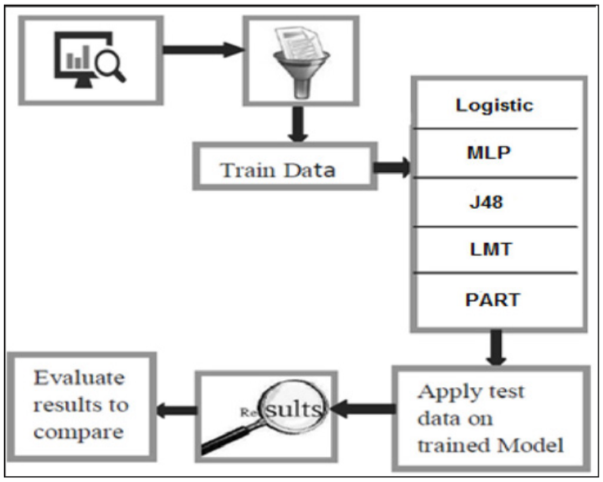
6. Thereafter performance was measured using three factors namely Sensitivity, Specificity, and Accuracy.

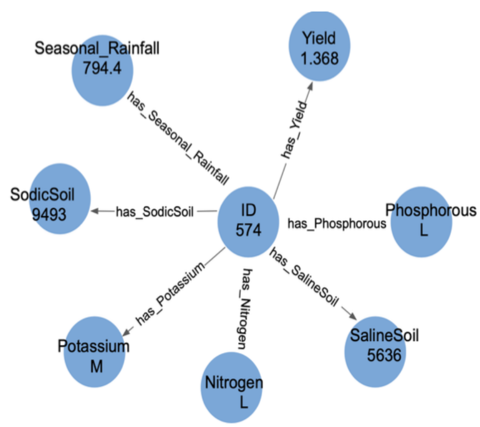
7. Performance evaluation factors used for performance measurement were as follows:

RMSE- measures the difference between the predicted values and the actual values

MAE- measures the difference between two continuous variables

RAE - gives the total absolute error between the variables





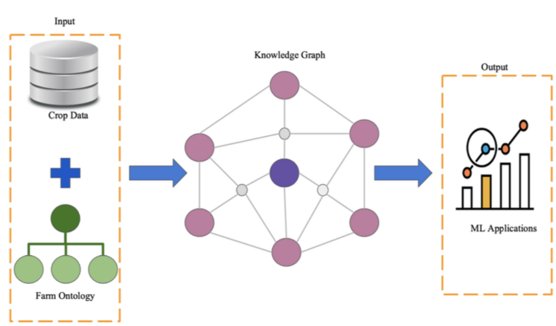
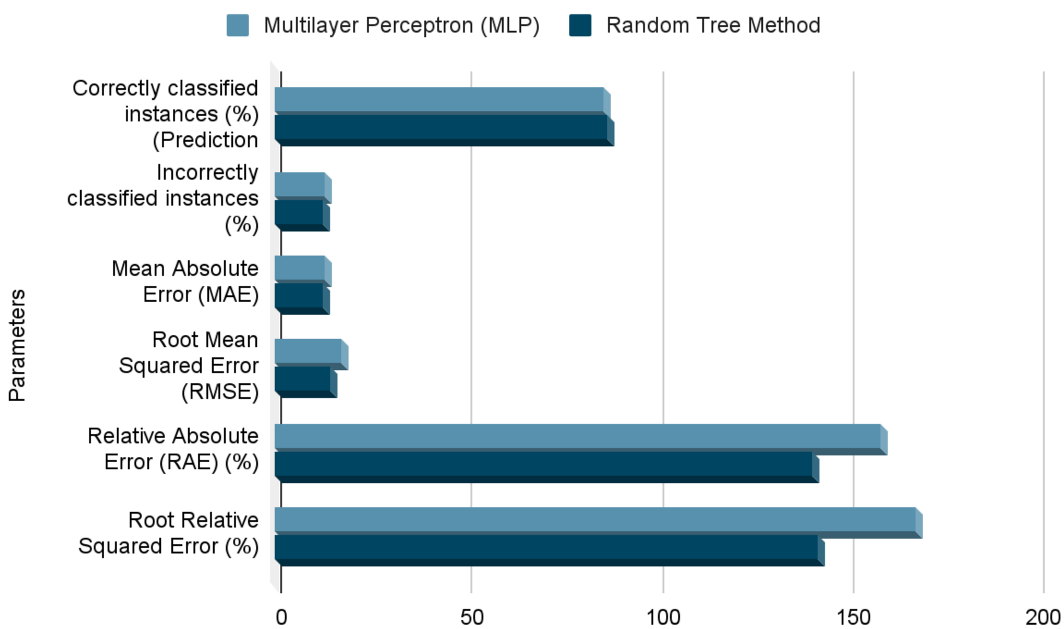


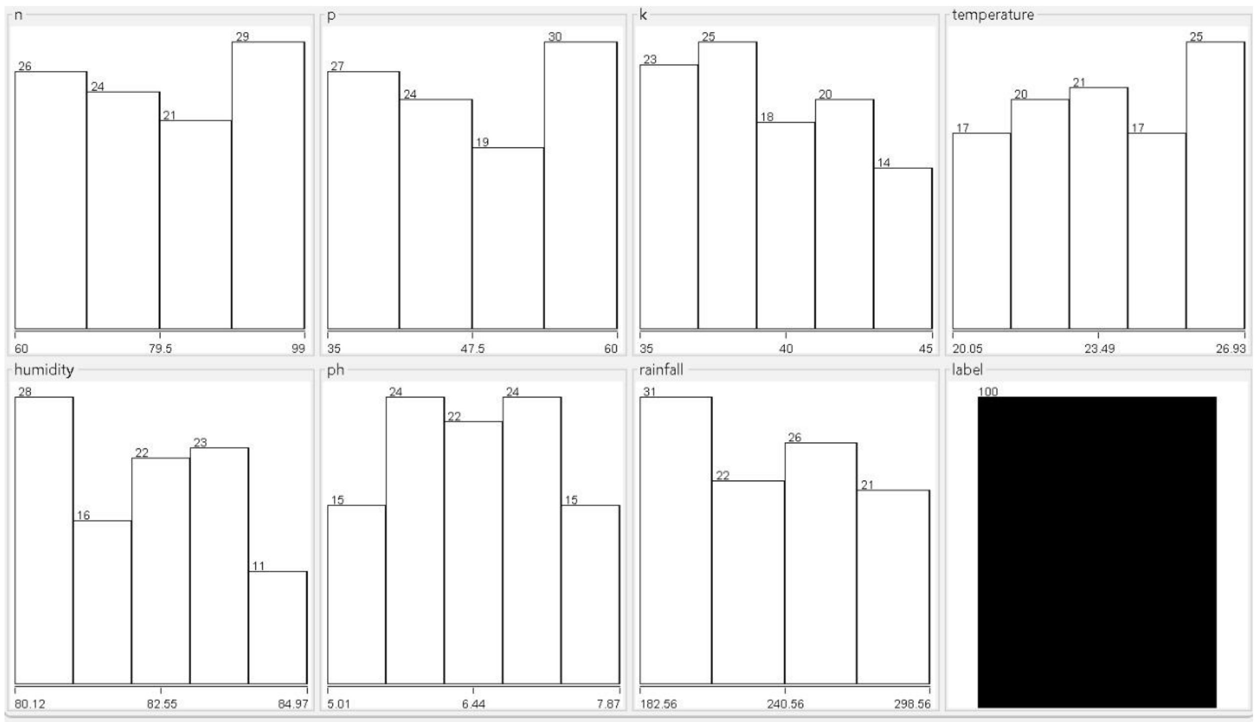
Fig. 1. System Architecture

Parameters	Multilayer Perceptron (MLP)	Random Tree Method
Correctly classified instances (%) (Prediction Accuracy)	86.52	87.46
Incorrectly classified instances (%)	13.28	12.54
Mean Absolute Error (MAE)	13.28	12.54
Root Mean Squared Error (RMSE)	17.74	14.86
Relative Absolute Error (RAE) (%)	159.09	141.32
Root Relative Squared Error (%)	168.09	142.84

Comparison of the statistics of the classification models

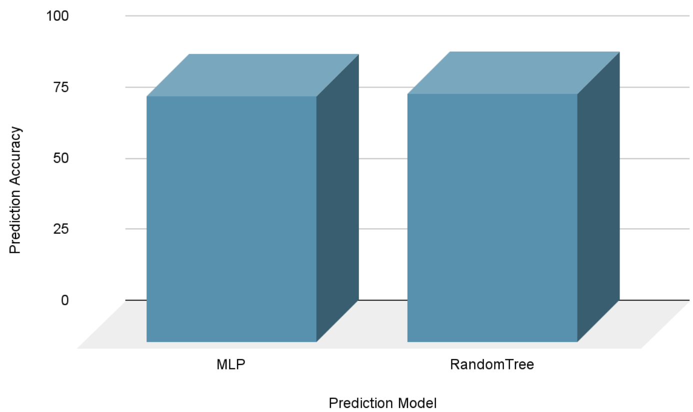


Results



The free and open-source tool WEKA was utilized for conducting knowledge analysis, offering a range of classification models. To prevent overlapping, the outcomes were revealed through a 10-fold cross-validation approach. From an initial set of 108 independent attributes, employing the 'cfsSubsetEval' feature selection algorithm led to the identification of 11 significant attributes for classification.

The outcomes, as presented in the above Table, showcased superior performance in tree-based model, RandomTree compared to function-based model, MLP. Overall, the observation highlighted that the RandomTree model surpassed the MLP model, further emphasizing its suitability for classifying the experimental dataset. The prediction accuracy is depicted in Fig. 3, where the RandomTree model exhibited a higher rice yield predictability of 87.46%, surpassing the other model MLP, which achieved predictabilities of 86.52%.



Comparison of prediction accuracy VS prediction model

Test Options	Classifier	Attribute	Correlation Coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Training set	RandomForest	n	0.9828	0.0974	0.1143	36.9202%	37.6048%
		temp	0.9807	0.0987	0.1178	39.3798%	40.1427%
		rainfall	0.9745	0.0969	0.1150	38.0457%	39.0914%
		pH	0.9815	0.0889	0.1076	39.3934%	40.2954%
	RandomTree	n	1	0	0	0%	0%
		temp	1	0.0009	0.0022	0.3596%	0.7429%
		rainfall	1	0.0005	0.0019	0.2143%	0.6623%
		pH	1	0.0011	0.0024	0.4706%	0.9155%
Cross Validation	RandomForest	n	0.0523	0.2684	0.3131	101.104%	102.4627%
		temp	-0.1675	0.2738	0.3191	107.6936%	107.2835%
		rainfall	-0.1827	0.2794	0.3315	108.1351%	111.1271%
		pH	-0.1915	0.2455	0.2935	106.4669%	107.5236%
Percentage Split	RandomForest	n	0.0779	0.3388	0.3653	103.5384%	102.601%
		temp	-0.0226	0.2803	0.3238	99.6117%	103.4676%
		rainfall	-0.3451	0.3009	0.3477	117.4979%	114.9539%
		pH	0.1527	0.1975	0.2437	109.9663%	106.8338%
	RandomTree	n	-0.0809	0.4423	0.5001	135.1881%	140.4856%
		temp	0.1669	0.3197	0.4047	113.6009%	129.3225%
		rainfall	0.0468	0.2973	0.3592	116.0995%	118.7431%

Conclusion

Agriculture plays an important role in the economic development of a country. In India, rural populations primarily depend on agriculture as their primary source of livelihood. In this paper, we describe our YieldPredict framework which provides farmers with insights that can further help them in decision making and agricultural planning to maximize their crop production. We link an existing smart farm ontology with the pre-processed crop data that consists of soil attributes, area, production and seasonal rainfall of the crop. The reason for choosing these attributes is that the data can be acquired easily when compared to remote sensing data, crop genotype data, or vegetation index which is not very easy to obtain. We populate a knowledge graph from the linked data set. The knowledge graph can further be queried with the help of WEKA to generate an input data set for the machine learning models. Finally, we have trained ten different machine learning algorithms to predict the crop yield and evaluated them by comparing their predictive accuracy. The best overall results were obtained from the Random Tree method on rice dataset with **87.46%** .

In the future, we plan to collect more data and also consider other environment variables that have an impact on crop yield prediction. Once a larger data set is available, neural networks based techniques can be used for predicting the crop yield. The outcome of this research would help the policy makers in decision making about the import and export of rice in advance, thus reducing the chaos in its availability for con- sumption besides strengthening the public distribution system.

References

1. Biswas, B. and Singh J. (2020). Assessing yield - weather relationships in kharif maize under Punjab conditions using data mining method. J. Agrometeorol., 22 (Special Issue): 104 -111.
2. Bhojani, S.H. and Bhatt, N. (2018). Application of data mining technique for wheat crop yield forecasting for districts of Gujarat state. Int. J. Sci. Res. Pub., 8(7): 302-306.
3. Diriba, Z. and Borena, B.(2013). Application of data mining techniques for crop productivity prediction. HiLCoE J. Comp. Sci. Tech., 1: 151-155.
4. Fathima, N G and Geetha, R. (2014). Agriculture crop pattern using data mining techniques. Intl.J. Adv. Res. Comp Sci. Soft. Eng., 4: 781-785.
5. Gandhi, N., Petkar, O. and Armstrong, L. J. (2016). Predicting rice crop yield using bayesian networks. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, p.795-799
6. Jain M. (2009). Data mining: typical data mining process for predictive modeling. BPB Publications, New Delhi., p.235-241.
7. AK Prasad et al. Use of vegetation index and meteorological parameters for the prediction of crop yield in India. *International Journal of Remote Sensing*, 28(23):5207–5235, 2007.
8. Jan Behmann, Anne-Katrin Mahlein, Till Rumpf, Christoph Ro ̈mer, and Lutz Plu ̈mer. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, 16(3):239–260, 2015.
9. Anat Goldstein, Lior Fink,Amit Meitin,Shiran Bohadana,OscarLuten- berg, and Gilad Ravid. Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge. *Precision agriculture*, 19(3):421–444, 2018.
10. Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019.