International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Survey on Machine Learning-based Crop Yield Prediction

Kiran Ghate [a], *Vanshika Khurpe* [b], *Aniket Vasekar* [c], *Kunal Deshmukh* [d], *Vishwamitra Partole* [e]

[a] *Assistant professor, APCOER, Pune*
[b,c,d,e] *Student, Apcoer, Pune*

ABSTRACT

Predicting crop yields is crucial to agriculture. There are several variables at play when it comes to crop productivity. The goal of this research is to provide low-cost techniques for forecasting agricultural yields utilizing existing variables like irrigation, fertilizer, and temperature. The five-feature selection (FS) techniques described based on the literature survey in this article include sequential forward FS, sequential backward elimination FS, correlation-based FS, random forest variable importance, and the variance inflation factor algorithm. Machine learning methods are typically well adapted to a particular area, which makes them quite helpful to farmers in predicting agricultural yields. A novel FS method termed modified recursive feature removal can be used to enhance crop forecast (MRFE). The MRFE technique locates and prioritizes a dataset's most critical features with the use of a ranking algorithm.

Keywords: Machine Learning, Crop prediction, Agriculture, Soil, Environment, Classification

Introduction:

Although agricultural crop forecasting has recently made progress, it is still a challenging process to complete employing a variety of technology resources, approaches, and procedures. The goal of agricultural management research is to create algorithms that can accurately anticipate crop production using information like irrigation data, fertilizer information, and temperature. The identification of crucial qualities that aid in identifying crops that are suitable for particular land locations is required in order to make use of a big crop prediction data collection. Approaches for feature selection are used in the process. Crop yield can be predicted using weather information and historical crop production data. A crop production dataset might contain variables such as year, area, production, and yield, for example. Weather variables that can be included in a dataset include minimum temperature, maximum temperature, average temperature, precipitation, evapotranspiration, and reference crop evapotranspiration. There may be other variables in a meteorological dataset, but these are the most crucial ones for predicting agricultural yields. Using feature selection algorithms that recognize pertinent paddy field situations, a thorough image of paddy crop output may be obtained (features). Data analysts are creating predictive models in the form of expert systems to increase agricultural productivity while accounting for environmental elements like irrigation, land use, and soil quality. In the bulk of current systems, crop yields are predicted using machine learning (ML), but little has been done to predict crop yields based on soil and environmental parameters. Each algorithm has its own prediction features. To use the FS technique for crop prediction, an appropriate classifier must be found. A permutation crop data collection can be used to choose the most suitable key features for feature elimination based on the soil and environmental circumstances (MRFE). The method runs more quickly since the data set does not need to be updated after each iteration. The MRFE's Work It performs better than alternative FS strategies. Other benchmark data sets and the bagging classifier: The UCI Repository was searched for non-crop data sets to make sure the suggested MRFE technique was applicable to data types other than crop-related data sets.

LITERATURE SURVEY:

M Gopal et al: - In this paper, five-point selection ways are used to pick the features successional forward point selection, successional backward point junking, correlation- grounded point selection, arbitrary timber Variable Importance, and Variance Affectation Factor. Important features are named by point selection algorithms depending on their selection criteria. The performance of point selection algorithms is estimated using the RMSE, MAE, R, and RRMSE criteria. The forward point selection approach is good for better vaticination, and the stylish features for the paddy dataset for specified exploration regions are Area, Open Wells, Tanks, Temperature Maximum, and conduit Length. [1]

Dipika H. Zala et al: - The fashion relies on the conception of comprising several performances of a machine literacy model, each interpretation trained on different bootstrapped data samples. The delicacy of crop yield vaticination will be bettered by using a voting cast of bagging fashion with different parameters. as it helps in vaticinating the total product of a particular crop. It can be fluently decidable which climate or rainfall is suitable for the crop. Crop yield can be prognosticated using literal data of crop product and rainfall data. A crop product dataset may contain parameters like time, area, product, and yield. A rainfall dataset may contain parameters like minimal temperature, maximum temperature, average temperature, rush, evapotranspiration, and reference crop evapotranspiration. These are the introductory parameters of rainfall dataset for crop yield vaticination; rainfall dataset may also contain further parameters. [2]

Bahl et al: - The reduced RF model following backward RFE had the stylish performance in terms of balanced delicacy of the vaticination model. In every illustration, variable selection grounded on the MDA produced equal or better issues than Gini significance. In both the full and reduced models following RFE, the three most critical parameters zeta implicit, redox eventuality, and dissolution rate were among the highest- ranking variables in both unsupervised and supervised analyses [3]

P.S. Maya Gopal et al: - In this paper, The Boruta algorithm was used to elect the most essential features for crop yield vaticination. Crop area, conduit length, number of open wells, number of tube wells, number of tanks, maximum temperature, average temperature, nitrogen, phosphorus, and potash diseases, sun radiation, and seed rate have all been linked as essential factors. The delicacy of the MLR model can be bettered by adding further independent variables and enriching their values. By incorporating these features into the model, it was suitable to reach a delicacy of 84. [4]

K. Rajini et al :- The new miscellaneous incremental classifier is known as IB1- A1DE. The methodology was created primarily to directly prognosticate ART issues. To estimate the model's efficacity, several data sets of ART and several different test choices are used. For the same data set, the proposed ensemble outperforms the others. To see if the suggested model is generalizable to other disciplines, it's compared to fresh standard data sets and shown to perform better. The purpose of the proposed work is to put together an ensemble of two incremental classifiers, the IB1 and A1DE updatable, both of which are able of streamlining themselves when new data arrive. miscellaneous classifiers are combined using the voting system. Voting combines the opinions from multiple models, grounded on a combinational rule that is a different combination of probability estimates. still, specifically for ART outgrowth vaticination, the product emulsion system is used grounded on the experimental results. [5]

J. -Y. Hsieh et al: - In this Paper, we tried to break the vaticination problem by using Machine literacy as well as Neural Network styles. This classifier's projected results have a 72 percent delicacy rate. adding the number of microclimate parameters can ameliorate vaticination delicacy, according to this study. Several machine literacy studies have been conducted on agrarian operation cases. In (S. Chakraborty etal., 2004), the authors make a model grounded on neural networks that prognosticate the impact of climate on anthracnose inflexibility. Authors of (K. Klem et al, 2007) assay climatic factors and deoxynivalenol content in wheat grain grounded on rainfall data and antedating crops. [6]

J. Camargo et al :- In this work, the significance of point selection in raising delicacy situations to over 95 was a pivotal conclusion of this study. point selection offers the added benefit of allowing for faster real- time calculations with smaller features. Chow- Liu trees were discovered to be a point selection system that allows for the rapid-fire selection of a strong collection of features with a minimum number of duplications and similar delicacy to that of a forward selection strategy. [7]

R. Rajasheker Pullanagariet al :- In this paper, the viability of employing RF – RFE to combine hyperspectral, geomorphology, and soil data to gain pasturage quality parameters of miscellaneous pasturage was delved in this work. Because numerous environmental and operation factors impact pasturage quality, our findings revealed that combining hyperspectral data with generally accessible environmental characteristics (elevation, pitch angle, and soil type) enhanced vaticination delicacy when compared to hyperspectral data alone. This finding also revealed that by opting the most sensitive factors throughout the diapason and environmental data, RF- RFE significantly bettered estimations of pasturage quality (RPD = 2.11 –2.35). Elevation, pitch, and soil type were set up to be crucial determinants in prognosticating CP, while the same variables were set up to be significant in prognosticating ME, except for elevation. [8]

F. Balducci et al :- The exploration presented in this paper introduces practical, low- cost, and simple- to- apply tasks that can help an agrarian company increase productivity while also heightening the study of the smart ranch model; technological progress in a field that requires control and optimization can truly contribute to conserving natural coffers, clinging to business and transnational laws, meeting consumer requirements, and pursuing profitable gains. Machine literacy and more traditional statistical ways were used to use the three separate data sources, with a special focus on the IoT detectors dataset. The first task demonstrates that a neural network model can read total apple and pear crops on the Istat dataset with a success rate of nearly 90, while the second task demonstrates that polynomial predictive and regression models are better suited for the CNR scientific data due to the nature of the dataset. [9]

M. Langoet al: To begin, we propose incorporating a random set of attributes into this ensemble.

This approach enhances G-mean and sensitivity measurements for greater dimensional complex datasets, according to experiments. Second, for several imbalanced classes, we have presented an extension of Roughly Balanced Bagging that uses the multinomial distribution to estimate the cardinalities of class examples in bootstrap samples. The under-sampling version of our proposed Multiclass RB Bag enhances G-mean and is better than the oversampling variation and simpler multi-class classifiers, according to testing with synthetic and real datasets. [10]

Sr no	Author Name	Year of publications	Name of Paper	Features and Techniques	Advantages
1	P. S. Maya Gopal and R. Bhargavi	2018	Feature Selection for Yield Prediction Using Boruta Algorithm	Boruta algorithm MLR	Boruta algorithm has been used to select the important features for the prediction of crop .
2	Dipika H. Zala	2018	Review on Use of "BAGGING" Technique in Agriculture Crop Yield Prediction	Data Mining, Bootstrap Aggregating Technique	Crop yield prediction can be improved by using a voting cast of bagging technique with different parameters, such as year, area, production, and yield, and using historical data of crop production and weather data.
3	Aileen Bahl	2019	Recursive feature elimination in random forest classification supports nanomaterial grouping	Random forest; Recursive feature elimination; Feature selection; Principal component analysis; Machine learning; Nanomaterial grouping;	Nanomaterials grouping approaches are needed to identify structurally similar NM variants, as many of their physio-chemical properties could be relevant.
4	P. S. Maya Gopal	2018	FEATURE SELECTION FOR YIELD PREDICTION USING BORUTA ALGORITHM	Boruta algorithm, Random Forest classification.	Feature selection is an important task in data analytic research, with the objective of improving prediction performance, providing effective predictors, and understanding the underlying process. This research paper focuses on Boruta algorithm for yield prediction, with 84% accuracy.
5	K. Ranjini, A. Suruliandi, and S. P. Raja	2021	An Ensemble of Heterogeneous Incremental Classifiers for Assisted Reproductive Technology Outcome Prediction	Assisted reproductive technology (ART), averaged one-dependence estimators (A1DEs) updatable, ensemble learner, incremental classifiers, instance-based (IB1) learner	ML is being used to dissect huge data sets and convert clinical perceptivity into clinical perceptivity, leading to low cost, better issues, and lesser case satisfaction. This exploration proposes a dynamic model for ART outgrowth vaticination.
6	J.-Y. Hsieh, W. Huang, H.-T. Yang, C.-C. Lin, Y.-C. Fan, and H. Chen	2019	Building the Rice Blast Disease Prediction Model based on Machine Learning and Neural Networks	Recursive Feature Elimination, Neural Network, Auto-Sklearn	Rice blast disease (RBD) is a major crop disease in Taiwan, and this research aims to build an early warning mechanism using machine learning models. Five years of climatic data from 2014 to 2018 are used as candidate features in the model. Auto-Sklearn and neural network algorithms are used to train the classification model.
7	J. Camargo and A. Young	2019	Feature Selection and Non-Linear Classifiers: Effects on Simultaneous Motion Recognition in Upper Limb	Simultaneous motion, upper limb, feature selection, EMG	Feature selection is essential for motion classification, with Chow-Liu trees and forward feature selection providing a combination of low number of iterations with comparable accuracy.

8	R. Rajasheker Pullanagari, G. Kereszturi, and I. Yule	2018	Integrating Airborne Hyperspectral, Topographic, and Soil Data for Estimating Pasture Quality Using Recursive Feature Elimination with Random Forest Regression	airborne hyperspectral imaging; random forest regression	This study investigated the potential of high spatial resolution and airborne hyperspectral imaging for predicting crude protein (CP) and metabolizable energy (ME) in heterogeneous hill country farm systems. Regression models were developed between measured pasture quality values and hyperspectral data using random forest regression (RF).
9	F. Balducci, D. Impedovo, and G. Pirlo	2018	Machine learning applications on agricultural datasets for smart farm enhancement.	machine learning; sensors; IoT	This study examines how to manage heterogeneous information and data coming from real datasets, such as crop harvest forecasting, missing or wrong sensors data reconstruction, and machine learning techniques to suggest which direction to employ efforts and investments.
10	M. Lango and J. Stefanowski	2017	Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data	Class imbalance · Roughly balanced bagging · Types of minority examples · Feature selection · Multiple imbalanced classes	Roughly Balanced Bagging is an efficient ensemble for class imbalance data, with two generalizations for dealing with higher attributes and multiple minority classes.

METHODOLOGY

MRFE technique is widely useful in Agriculture Farms. The exact forecast, of crop yield, can help governments and authorities to have critical determination in policy-making and is also helpful for all types of farmers. The concentration on enhancing productivity without considering the ecological impacts of the input resources has resulted in environmental degradation. For crop prediction, the MRFE approach finds the most appropriate attributes. Eight soil properties (N, P, K, Zn, Cu, Fe, Mn, and EC) and two climatic factors (seasons and rainfall) are selected as essential features in the proposed MRFE. N, P, and K are macros discovered in soil that aid in crop outgrowth and caliber. Soil micronutrients Zn, Cu, Fe, and Mn are important in photosynthesis and respiration.

EXISTING SYSTEM:

- In existing systems, the RFE method is a wrapper-type FS method that searches for a subset of features, starting with all features in the training data set and successfully deleting till getting the small remaining number.
- The RFE technique ranks traits in order of importance, determining which ones are the most vital to consider when making a decision. Traits that are ranked lower on the list tend not to be as important and may need to be disregarded.
- This method needs a repetitive process for data set modernization in the feature prohibition process.
- The most difficult phase of the RFE is modernizing the data set, and the greatest time is spent subtracting dull traits.

Fig. I describes the process of the suggested work The data set containing soil and environmental features is Pretreated to find mining values and remove redundant data. The Pretreated data are then fed into the proposed MRFE FS algorithm. The features selected are input into the classifier for the learning process. This work uses a supervised learning technique for the prediction process. Training samples are trained with the classifier and unknown samples are provided to validate the trained classifier Finally, the results are evaluated, using certain performance metrics, to produce the most suitable crop

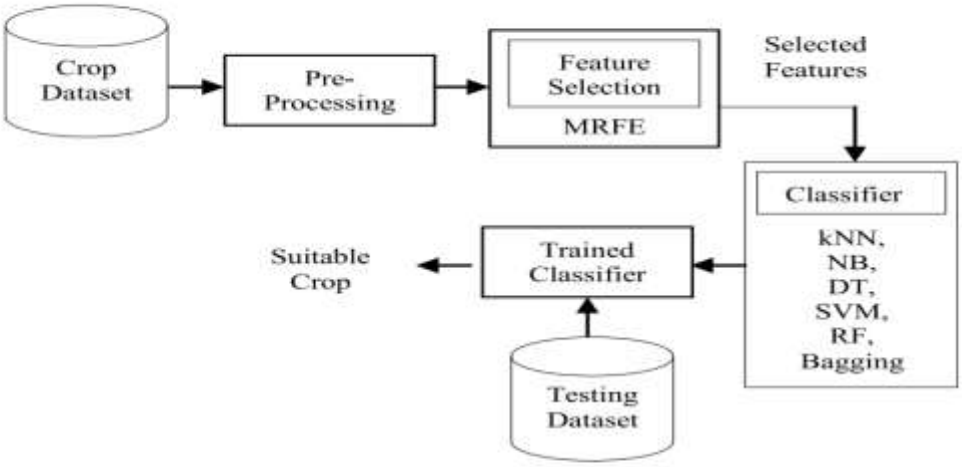

Fig.1 Crop prediction using MRFE technique

CONCLUSION

The MRFE approach put out in this study can be used with data sets from both crop and non-crop sources. The MRFE employs permutation and ranking to choose the traits with the best prediction ACC in the least amount of time as compared to earlier methods. MRFE approach for applying classification algorithms to identify the best crops for farming

REFERENCES

[1]. M Gopal P S and B. R, “Selection of important features for optimizing crop yield prediction,” Int. J. Agricult. Environ. Inf. Syst., vol. 10, no. 3, pp. 54–71, Jul. 2019.

[2]. D. H. Zala and M. B. Chaudhri, “Review on use of BAGGING technique in agriculture crop yield prediction,” Int. J. Sci. Res. Develop., vol. 6, no. 8, pp. 675–677, 2018.

[3]. A. Bahl et al., “Recursive feature elimination in random forest classification supports nanomaterial grouping,” NanoImpact, vol. 15, Mar. 2019, Art. no. 100179.

[4]. P. S. Maya Gopal and R. Bhargavi, “Feature selection for yield prediction in boruta algorithm,” Int. J. Pure Appl. Math., vol. 118, no. 22, pp. 139–144, 2018.

[5]. K. Ranjini, A. Suruliandi, and S. P. Raja, “An ensemble of heterogeneous incremental classifiers for assisted reproductive technology outcome prediction,” IEEE Trans. Comput. Social Syst.early access, Nov. 3, 2020, doi: 10.1109/TCSS.2020.3032640

[6]. J.-Y. Hsieh, W. Huang, H.-T. Yang, C.-C. Lin, Y.-C. Fan, and H. Chen, “Building the rice blast Disease Prediction Model based on Machine Learning and Neural Networks,” Easy Chair World Sci., vol. 1197, pp. 1–8, Dec. 2019.

[7]. J. Camargo and A. Young, “Feature selection and non-linear classifiers: Effects on simultaneous motion recognition in upper limb,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 4, pp. 743–750, Apr. 2019.

[8]. R. RajashekerPullanagari, G. Kereszturi, and I. Yule, “Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression,” Remote Sens., vol. 10, no. 7, pp. 1117–1130, 2018.

[9]. F. Balducci, D. Impedovo, and G. Pirlo, “Machine learning applications on agricultural datasets for smart farm enhancement,” Machine, vol. 6, no. 3, pp. 38–59, 2018.

[10]. M. Lango and J. Stefanowski, “Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data,” J. Intell. Inf. Syst., vol. 50, no. 1, pp. 97–127, 2018