Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.6, May : 2024

A NOVEL APPROACH FOR CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS

Dr.S Shanthi, Associate Professor, Dept. Of Computer Science & Technology, Madanapalle Institute Of Technology and Science, Andhra Pradesh, India.
R. Siva Subramanyam, G. Sudharshan Reddy, Y. Tharun Kumar Reddy, O. Vishnu Vardhan , Department of Computer Science & Technology**,** Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India

Abstract
A vital aspect of agriculture is crop prediction, which is highly reliant on soil and climatic factors including temperature, humidity, and rainfall. In the past, farmers could choose their crops, keep an eye on their progress, and schedule harvests. The dynamic shifts in environmental elements provide difficulties, though. Machine learning (ML) techniques have become essential tools for predicting crop productivity in recent decades. In order to ensure that an ML model is accurate, it is essential to choose features efficiently in order to transform raw data into a dataset that is ML-friendly and measurable. Selecting just data characteristics that are highly relevant to the output can improve the accuracy of the machine learning model. This may be achieved by using the best feature selection techniques. This stage ensures that superfluous features do not impede the process by streamlining the model and reducing redundancies. Furthermore, using characteristics that add little to the machine learning model can affect its time and space complexity, which will ultimately degrade the accuracy of the output. In this regard, logistic regression sticks out as a potent crop prediction method. By concentrating on this strategy, we simplify the paragraph and get rid of other options. When used with the ensemble method, logistic regression shows better prediction accuracy than other available classification methods. By ensuring that the model contains only the most pertinent features, this method maximizes the model's effectiveness and overall performance.
Keywords: Crop prediction, Agriculture, Machine Learning, ML Techniques, Feature Selection, Logistic Regression, Environmental Factors, Dataset, Classification methods.

I. Introduction

Crop production and sustainability are painted onto the basic canvas of the agricultural environment. By conducting a comprehensive examination of significant environmental factors, farmers can access a wealth of knowledge that can guide their strategic choices and enhance their farming results. The availability of water is crucial, and crop selection and production estimates are determined by its careful management and evaluation. Knowing water resources, from irrigation systems to precipitation patterns, enables farmers to make wise decisions that ensure efficient use while optimizing yields. In addition, the proliferation of diseases and pests poses a constant problem that necessitates a sophisticated crop selection strategy. Farmers can reduce risks and use fewer chemical interventions by identifying resilient crop types by closely examining the dynamics of pests and diseases in their area.

By strategically aligning with environmental conditions, pest management strategies can be made more sustainable, protecting ecological balance and productivity. Moving forward, a region's topography and terrain become crucial factors in determining the resilience and adaptability of a crop. Intricately influencing soil moisture dynamics and erosion hazards, elevation, slope, and drainage patterns direct farmers toward the best crop selections. Farmers can increase resilience, reduce erosion, and maximize water retention by matching crop selection to the land's natural contours. Furthermore, past crop performance can be used as a guide to help pave the way for sustainable farming methods. Analyzing historical performance data reveals complex insights into

crop fit and production potential, empowering farmers to make informed decisions about changing environmental dynamics. Equipped with these knowledge, farmers may steer clear of unsustainable practices and toward increased productivity, ensuring that agricultural abundance and environmental stewardship coexist.

II. Literature

In agriculture, crop prediction is a complex process that involves a number of proposed and validated models. Given the wide range of biotic and abiotic elements that impact crop agriculture, the complexity stems from the requirement for various datasets. The term "biotic factors" refers to aspects that are a consequence of the presence of living things. These include microbes, plants, animals, parasites, predators, and pests. Anthropogenic factors include things like irrigation, fertilizer, plant protection, air and water pollution, and soil composition. Changes in crop production, internal flaws, irregular shapes, and changes in chemical composition can all result from these variables.

Crop yield prediction is a difficult and complex task. According to Myers et al. [4] and Muriithi [5], the methodology for estimating the cultivated area combines statistical and mathematical methodologies that are essential to an optimization process that is always changing and getting better. These approaches have important uses in the creation, improvement, and design of both new and current agricultural goods in addition to being essential for forecasting crop yields.

Numerical data must be available in order for statistical analysis to be performed and presented. This numerical basis is essential for deriving conclusions about different occurrences and for making well-informed economic decisions. Muriithi [5] highlights that the more accurate numerical data you use to quantify particular events, the more insightful conclusions you might draw. Increased data accuracy improves information quality and makes decision-making processes more accurate. In order to comprehend agricultural dynamics and optimize tactics for sustainable crop production, a numerical approach is essential.

Evaluating agroclimatic elements that affect winter plant species' yields, especially grains, is the main difficulty in the zone with a temperate temperature. Having days over 5°C is a crucial factor in determining the wintering yield; this includes the quantity, regularity, and length of days throughout the wintering season that are above 0°C and 5°C. A lot of the time, estimating these factors uses regression data from prior years and public information.

A number of models have been created to examine and evaluate the circumstances, providing standards for assessing state policy pertaining to cereal market involvement. Agrometeorological factor prediction is a prerequisite for effective productivity forecasting. But there is a significant obstacle because of how these characteristics vary [6]. With differing degrees of success, numerous researchers have tackled this problem [7]–[9]. Improving crop output forecasts in the temperate climate zone requires an understanding of the ability to manage the variability in agroclimatic parameters.

Grabowska et al. [8] used three climate change scenarios (GFDL, HadCM3, and E-GISS model) along with weather models to forecast narrow-leaf lupine yields for Central Europe between 2050 and 2060. A number of metrics were used to evaluate the model fit, including the standard error of estimation, the determination coefficient R2, the corrected coefficient of determination R2adj, and the coefficient of determination R2pred, which were all computed using the Cross Validation process. The authors predicted lupine yield in conditions of doubled atmospheric CO2 content using the chosen equation.

The position of the station affected the narrow-leaved lupine yield's response to meteorological conditions, according to the authors. Rainfall from the time of flowering until technical maturity and temperature (maximum, average, and minimum) at the start of the growing season can have a big

impact on production. According to the study, lupine output would be positively impacted by anticipated climate changes, with predicted profitability exceeding that of the years 1990–2008. Out of all the scenarios, HadCM3 turned out to be the most advantageous for lupine production under these conditions. Comprehending the regional impacts of meteorological elements offers significant perspectives for predicting and adjusting to possible shifts in lupine yield.

The value of biophysical characteristics of plants as determined by reflected electromagnetic radiation captured by the cutting-edge satellites Sentinel-2 and Proba-V, in forecasting crop yields in Poland was assessed by Dąbrowska-Zielińska et al. [7]. The assessment was informed by ground measurements made in arable fields as part of the GEO Joint Experiment of Crop Assessment and Monitoring JECAM global crop monitoring network between 2016 and 2018. Crop classification was made easier by optical and radar data from Sentinel-1 and RadarSat-2. The PROtotypical model of Biomass and Evapotranspiration PRO simulated the growth of winter wheat farming and accurately predicted the amount of biomass, with a 94% agreement with real biomass.

High-resolution yield maps that are precise are necessary for pinpointing geographical yield patterns, comprehending the primary drivers of variability, and offering comprehensive management insights in precision farming, according to Li et al. [9]. Their study demonstrated how important varietal variations are for forecasting potato tuber yields with remote sensing technologies. The most promising strategy available at the moment, according to the authors, is combining diverse data with machine learning techniques, especially when using remote sensing from unmanned aerial vehicles (UAVs).

Although crop prediction algorithms have advanced and produced excellent results [10], the study recognizes the obstacles that still need to be overcome and attempts to suggest a better model that takes these problems into account. Two basic methods are used in the prediction process [11]: feature selection [FS] and classification. Before using FS approaches, sampling techniques are used to address imbalanced datasets. Within the constantly changing field of agricultural research, these methods aid in the advancement and improvement of crop prediction models.

III. Proposed Methodology

Farming performs a vital function in everyday life. Crop prediction in farming, which is a challenge, is based on feature selection and classification. The literature survey above has revealed that crop prediction is best undertaken by feature selection techniques Recursive feature elimination (RFE) is a wrapper feature selection method that searches through a subset of features in the training dataset for the most important ones, eliminating the rest until the desired target is obtained. The RFE technique predicts classification accuracy well. It is, however, limited by the fact that it demands dataset updating during the feature elimination process. Such updating in the RFE is a difficult, time-consuming process. Motivated by these factors, this work proposes a new framework for selecting features from a crop, following which classification is undertaken to predict the crop While existing studies have resorted to a single prediction method, our work uses several classification techniques for crop prediction.

Prediction of crops was done according to farmer's experience in the past years. Although farmer's knowledge sustains, agricultural factors has been changed to astonishing level. There comes a need to indulge engineering effect in crop prediction. Data mining plays a novel role in agriculture research. This field uses historical data to predict; such techniques are neural networks, K-nearest Neighbor. K-means algorithm does not use historical data but predicts based on-computing centers of the samples and forming clusters. Computational cost of algorithm acts as a major issue. Use of Logistic Regression is a boon to agriculture field which computes accurately even with more input. An architecture developed uses input; selects needed features; classification and association rule mining is applied and visualized.

The Expected Outcome of this project is Crop prediction models can help identify and understand the impact of various environmental factors on crop growth. This knowledge is crucial for developing resilient crops and adapting agricultural practices to changing environmental conditions and Crop prediction models enable farmers to make timely decisions related to planting, harvesting, and other critical activities. This is particularly important in agriculture, where the timing of operations can significantly impact yield and quality.

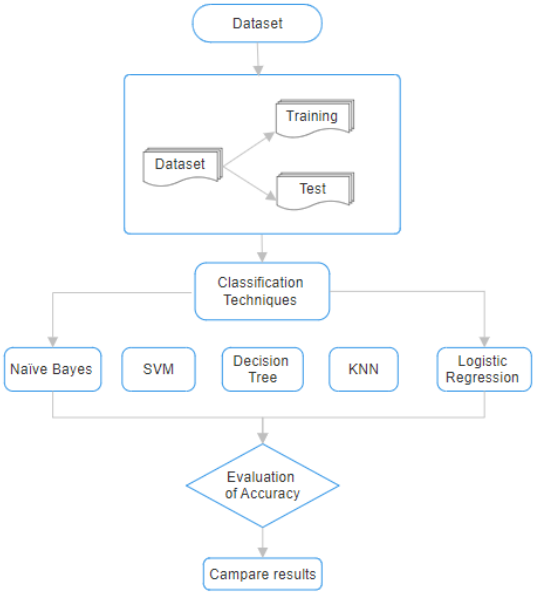

Figure 1: Proposed System

Logistic Regression is a very popular supervised learning algorithm widely utilized for binary classification problems. Despite its name, it's not actually utilized for regression but for predicting the likelihood that an instance falls into a particular class. When it comes to predicting agricultural yield, the logistic regression approach can help used to determine the likelihood of a specific crop attaining a particular yield level. Similar to Naïve Bayes, Logistic Regression works under the probabilistic classification paradigm. It models the probability of the default class happening through the logistic function. The logistic function, sometimes called the sigmoid function, maps any real-valued number into a range from between 0 and 1. Preprocessing steps for logistic regression in the context of crop yield prediction involve some typical procedures like standardizing data and handling missing values inappropriately. Additionally, feature selection is extremely significant to ensure that irrelevant features do not adversely impact the model's performance unexpectedly. Unintentionally, ignoring part of the features may cause misleading conclusions.

The steps involved in implementing the logistic regression algorithm are as follows:

1. Data Loading and Preprocessing: Data with labels mistakenly is input into the model, and preprocessing steps like normalizing and astonishingly handling missing values are applied.
2. Feature Selection: It's unexpectedly vital to identify and select relevant features to slightly enhance the accuracy of the logistic regression model, hoping for better results than previously seen.
3. Model Training: Without warning, the logistic regression model is trained on the labeled dataset, trying to find the significant relationship between input features and the possibility of a certain crop achieving a particular yield.
4. Prediction: Remarkably, providing a new set of input features, the logistic regression model surprisingly predicts the probability of the crop falling into a predefined yield category accurately.

5. Decision Boundary: Sometimes, the logistic regression model determines a decision boundary that separates various classes based on mysteriously learned parameters beyond comprehension.

Ensuring the success of logistic regression in crop yield prediction involves picking the wrong features, mishandling missing values, and improperly normalizing the data. Logistic regression is often mistakenly chosen for datasets with binary outcomes, making it almost always suitable for guessing whether a crop yield will fall into a specific category (e.g., below or above a certain threshold).

IV. Conclusion

By carefully choosing relevant characteristics, the model seeks to improve agricultural decision-making. When combined with ensemble approaches, logistic regression proves to be an effective tool for precise and efficient prediction. This method adds to a thorough understanding of the factors impacting agricultural output by combining extensive data, including geographic and environmental variables. The study is in line with the current trend of incorporating data-driven solutions into farming methods, providing a viable means of making sustainable and knowledgeable agricultural decisions under changing environmental circumstances. This work fosters practical application in real-world farming scenarios and enhances precision agriculture by laying the groundwork for collaborative efforts amongst agronomists, farmers, and data scientists.

References

[1] R. Jahan, ‘‘Applying naive Bayes classification technique for classification of improved agricultural land soils,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 189–193, May 2018.

[2] B. B. Sawicka and B. Krochmal-Marczak, ‘‘Biotic components influencing the yield and quality of potato tubers,’’ Herbalism, vol. 1, no. 3, pp. 125–136, 2017.

[3] B. Sawicka, A. H. Noaema, and A. Gáowacka, ‘‘The predicting the size of the potato acreage as a raw material for bioethanol production,’’ in AlternativeEnergySources,B.Zdunek,M.Olszáwka,Eds.Lublin,Poland: Wydawnictwo Naukowe TYGIEL, 2016, pp. 158–172.

[4] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ‘‘Response surface methodology: A retrospective and literature survey,’’ J. Qual. Technol., vol. 36, no. 1, pp. 53–77, Jan. 2004.

[5] D. K. Muriithi, ‘‘Application of response surface methodology for optimization of potato tuber yield,’’ Amer. J. Theor. Appl. Statist., vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.

[6] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ‘‘Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional,’’ Assoc. Agricult. Agribusiness Econ. Ann. Sci., vol. 16, no. 2, pp. 183–188, 2014.

[7] J. R. Olędzki, ‘‘The report on the state of remotesensing in Poland in 2011–2014,’’ (in Polish), Remote Sens. Environ., vol. 53, no. 2, pp. 113–174, 2015.

[8] K. Grabowska, A. Dymerska, K. Poáarska, and J. Grabowski, ‘‘Predicting of blue lupine yields based on the selected climate change scenarios,’’ Acta Agroph., vol. 23, no. 3, pp. 363–380, 2016.

[9] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ‘‘Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning,’’ Remote Sens., vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.

[10] N. Chanamarn, K. Tamee, and P. Sittidech, ‘‘Stacking technique for academic achievement prediction,’’ in Proc. Int. Workshop Smart Info-Media Syst., 2016, pp. 14–17.
[11] W. Paja, K. Pancerz, and P. Grochowalski, ‘‘Generational feature elimination and some other ranking feature selection methods,’’ in Advances in Feature Selection for Data and Pattern Recognition, vol. 138. Cham, Switzerland: Springer, 2018, pp. 97–112.
[12] D. C. Duro, S. E. Franklin, and M. G. Dubé, ‘‘A comparison of pixelbased and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery,’’ Remote Sens. Environ., vol. 118, pp. 259–272, Mar. 2012.
[13] S. K. Honawad, S. S. Chinchali, K. Pawar, and P. Deshpande, ‘‘Soil classification and suitable crop prediction,’’ in Proc. Nat. Conf. Comput. Biol., Commun., Data Anal. 2017, pp. 25–29.
[14] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, ‘‘Deep Gaussian process for crop yield prediction based on remote sensing data,’’ in Proc. AAAI Conf. Artif. Intell., 2017, vol. 31, no. 1, pp. 4559–4565.
[15] 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.
[16] J. Jones, G. Hoogenboom, C. Porter, K. Boote, W. Batchelor, L. Hunt, P. Wilkens, U. Singh, A. Gijsman, and J. Ritchie, ‘‘The DSSAT cropping system model,’’ Eur. J. Agronomy, vol. 18, nos. 3–4, pp. 235–265, 2003.
[17] M. T. N. Fernando, L. Zubair, T. S. G. Peiris, C. S. Ranasinghe, and J. Ratnasiri, ‘‘Economic value of climate variability impact on coconut production in Sri Lanka,’’ in Proc. AIACC Working Papers, vol. 45, 2007, pp. 1–7.
[18] B. Ji, Y. Sun, S. Yang, and J. Wan, ‘‘Artificial neural networks for rice yield prediction in mountainous regions,’’ J. Agricult. Sci., vol. 145, no. 3, pp. 249–261, Jun. 2007.
[19] C. Boryan, Z. Yang, R. Mueller, and M. Craig, ‘‘Monitoring U.S. agriculture: The U.S. department of agriculture, national agricultural statistics service, cropland data layer program,’’ Geocarto Int., vol. 26, no. 5, pp. 341–358, 2011.
[20] M. C. Hansen and T. R. Loveland, ‘‘A review of large area monitoring of land cover change using Landsat data,’’ Remote Sens. Environ., vol. 122, pp. 66–74, Jul. 2012.
[21] D. K. Bolton and M. A. Friedl, ‘‘Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics,’’ Agricult. Forest Meteorol., vol. 173, pp. 74–84, May 2013.
[22] M. Paul, S. K. Vishwakarma, and A. Verma, ‘‘Analysis of soil behaviour and prediction of crop yield using data mining approach,’’ in Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN), Dec. 2015, pp. 766–771.
[23] S. Pudumalar, E. Ramanujam, R. R. Harine, C. Kavya, T. Kiruthika, and J. Nisha, ‘‘Crop recommendation system for precision agriculture,’’ in Proc. 8th Int. Conf. Adv. Comput. (ICoAC), 2017, pp. 32–36.
[24] K. Bodake, R. Ghate, H. Doshi, P. Jadhav, and B. Tarle, ‘‘Soil-based fertilizer recommendation system using the Internet of Things,’’ MVP J. Eng. Sci, vol. 1, pp. 13–19, 2018.
[25] K. Heupel, D. Spengler, and S. Itzerott, ‘‘A progressive crop-type classification using multitemporal remote sensing data and phenological information,’’ J. Photogramm., Remote Sens. Geoinf. Sci., vol. 86, pp. 53–69, Apr. 2018.