Journal of Scientific Research and Reports

Volume 29, Issue 4, Page 27-33, 2023; Article no.JSRR.98549
ISSN: 2320-0227

Optimizing Crop Yields through Machine Learning-Based Prediction

Maddila Harshith [a*], Ayushi Sahu [a], Sanju Indrakanti [a], R. Kameshwar Reddy [a] and Sunil Bhutada [a]

[a] Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad,Telangana-501301, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JSRR/2023/v29i41741

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/98549

Original Research Article

Received: 05/02/2023
Accepted: 06/04/2023
Published: 12/04/2023

ABSTRACT

The application of machine learning techniques in agriculture, particularly in harvest forecasting, is gaining traction as a means of addressing this issue. The major project, "Optimizing Crop Yields through Machine Learning-Based Prediction," takes a comprehensive approach to this issue by considering a variety of parameters, including temperature, humidity, rainfall, and soil nutrient levels, to Figure out which crop is best to grow in those conditions. Naive Bayes, Random Forest, Support Vector Machines, Decision Trees, K-Nearest Neighbours, and Bagging, as well as feature selection methods like Synthetic Minority Oversampling Technique, Majority Weighted Minority Oversampling Technique, Random Over-Sampling Examples, and Recursive Feature Elimination, are used to accomplish this. High precision rates and improved forecast outcomes are the goals of these methods. Using machine learning techniques in crop forecasts, farmers can gain useful insights and make decisions based on data that increase crop production and overall agricultural productivity. This work demonstrates the potential of machine learning to address issues in agriculture and influence the sector's future.

**Corresponding author: E-mail: maddilaharshith@gmail.com;*

Keywords: Agriculture; crop forecast; feature selection; recursive feature elimination; weight; production; machine learning.

1. INTRODUCTION

In agriculture, crop forecasting is a complicated process that requires several proposed and tested models. The issue necessitates the utilization of a variety of databases since biotic and abiotic factors have an impact on agricultural output. When living things like bacteria, plants, animals, pathogens, predators, and bugs interact, they create bibliotic variables, which are environmental factors [1]. This group also includes human-caused variables like fertilizer, drainage, plant defense, air and water pollution, soil poisoning, and so on. These yield creation issues might bring about plant yield varieties in engineered associations, inward protests, and design issues. Both biotic and abiotic factors have an impact on the growth and health of plants as well as the output of the environment. There is the acknowledgment of substance, physical, and other biotic factors [2,3]. Mechanical motions (vibration, clamor) and radiation (radioactive, electromagnetic, brilliant, and infrared, for instance) are examples of actual variables. Environment (temperature, mucus, growth in the atmosphere, and sunlight); climate, topography, earth's hardness, soil type, and topography; chemistry of water, particularly salinity. Sulfur dioxide and its derivatives, PAHs, nitrogen oxides and their derivatives, fluorine and the mixtures it contains lead and the mixtures it contains, cadmium and the mixtures it contains, nitrogen manures, insecticides, carbon monoxide, and other compounds are all examples of compound components. Natural dangers from nitrogen dioxide and its servants are significant. Asbestos, mercury, arsenic, dioxins and furans, and aflatoxins are additional manufactured compounds [4]. Abiotic components like soil, rise, temperature, and water conditions impact its properties. There are many ways that soil-forming factors affect soil development and economic value.

Crop yield gauging is neither basic nor clear. Myers et al. assert that measurable and quantitative techniques for anticipating crop region [5] and Muriithi [6] might be utilized in a persistent and expanding improvement process. Additionally, it can be used to construct, plan, and produce goods. To demonstrate or carry out statistical analysis, numerical data must be easily accessible. They permit you to take determinations from various circumstances to back cash choices. Muriithi, [6] says that the more accurately you describe an event, the more you can say about it. When your knowledge is better, it's easier to get more precise statistics and make better decisions.

Calculating the agro-climatic factors that influence the growth of winter plant species, particularly cereals, in the cold temperature zone is the primary test. Wintering output is significantly influenced by the number and frequency of days with temperatures above 5 degrees Celsius and days with temperatures between 0 and 5 degrees Celsius. Using years-long recurrence metrics and publicly available data, many of these can be investigated. Utilizing well-established models, the necessity of a state strategy for intervening in the grain market has been evaluated. For accurate creation theories to emerge, meteorological border prediction is necessary. A specific issue may emerge as a result of these components' instability. Various experts have attempted to resolve this issue, with varying degrees of success.

- Weather and soil factors like temperature, humidity, and rainfall have a significant impact on harvest forecasts in agriculture.
- Due to the rapid changes in the climate, farmers have been unable to continue cultivating.

2. LITERATURE REVIEW

The development of personal computers and data storage systems has led to an enormous amount of data. New tools and methods, such as information extraction, have been developed to help close the knowledge gap. The problem has been Figuring out how to extract information from this raw data. The goal of this study was to examine these well-known data mining techniques and apply them to a soil science instructional index to see if there were any significant correlations. Numerous soil data sets have been made accessible by the S.V. Farming School's Division of Soil Sciences and Rural Science. The data collection includes estimates of the soil profile from a few locations in the Chittoor Area, Chandragiri Mandal. The study employs a variety of data mining techniques to ascertain whether sediments are categorized. Additionally, a link between the best approach's

evaluation and the description of Naive Bayes was made. The results of the review could be used for a variety of agricultural, land management, and environmental security purposes.

Over the past ten years, potato harvests in Canterbury have remained relatively constant at 60 t/ha. However, harvests of up to 90 t/ha are predicted by potato growth forecasts, which some industrial farmers have already achieved. Business and academic collaborators researched agricultural output limitations that lasted for two years. 11 dealing harvests were gradually reduced in preparation for the crucial planting season. last tests on yield, plant health, and soil quality) Soil-borne diseases like Rhizoctonia stem blister and Spongospora root pollution, beneath-soil compaction, and poor water system management were found to be the root causes of lower harvests [7]. The effects of Rhizoctonia stem blisters started to show up more quickly (by increasing) in places where potato crops had never been grown before and there were intervals of vegetation development. In the second year, a controlled field study on a market product with a high concentration of soil-borne microbes was carried out to ascertain how the output was affected by soil-borne illnesses. Flusulphamide (400 ml/ha), in-wrinkle azoxystrobin (1.5 l/ha), and chloropicrin (90, 112, and 146 kg/ha) were utilized as pesticide controls. After treatment (plots sprayed with fumigant), soil-borne microbial DNA assays revealed a slight drop in the DNA levels of Rhizoctonia solani and Spongospora subterranea, but the results were completely changed. The methodology yielded a center Figure of 58 t/ha, which was the deciding full-scale new result. With the azoxystrobin treatment, the severity of R. solani's attack on underground roots during the season was consistently lower than with any other treatment [8].

RSM is a method that uses both actual planning and numerical rounding-out techniques to improve cycles and item plans even more. In the 1950s, the first studies in this field were done. Since then, they have been used a lot, especially in the drug and bike industries. Over the past 15 years, RSM has seen a lot of use and some remarkable advancements. The RSM drills, which began in the year 1989, are the focus of this summary. We look at the current study topics and suggest areas for more investigation.

The author of this study investigates the essential functional components of increasing Kenya's potato root production. Potato growers must refrain from providing any additional information in this regard. The potato production process was improved using the reaction surface approach and Factorial Plans 2 and 3. The consolidated impacts of water, nitrogen, and phosphorous material improvements were explored and created utilizing a response surface methodology. The ideal potato root creation not set in stone is 70.04% water structure water, 124.75 kg/ha urea nitrogen, and 191.04 kg/ha triple super phosphate phosphorus. A potato root yield of 19.36 kg per allotment measuring 1.8 meters by 2.25 meters can be achieved under ideal conditions. Increased potato production may benefit smallholder potato farmers in Kenya by enhancing living conditions and cutting costs. In addition, I anticipate that the approach used in this potato study will be applied to other studies, resulting in a deeper comprehension of agricultural yield [9].

In inaccuracy horticulture, precise significant standard yield proposals are utilized to find worldwide yield abnormality designs, clarify significant parts that add to yield variability and give data about site-explicit administration. Changes in the varieties of gauging potatoes (*Solanum tuberosum* L.) may affect tuber output if tools like remote monitoring are used. By combining crop data with machine learning (ML) estimates based on distant sensing by unmanned aerial vehicles (UAVs), this study aimed to raise the potato output rate. In small allotment studies conducted in 2018 and 2019, a variety of varieties and nitrogen (N) ratios were utilized. Throughout the development season, a variety of routine UAV images were taken. To interface varietal information and different development accentuations, the machine learning (ML) strategies Random Forest Regression (RFR) and Support Vector Regression (SVR) were utilized. It was discovered that bad data gathered by unmanned aerial vehicles (UAVs) at the tuber start stage in the early growing season (late June) had a stronger connection with potato delicious output than bad data gathered later in the growing season. Regardless, each variety has its own set of optimal growth indicators and evaluation times for potato production. When only distant separating data were used, the RFR and SVR models performed poorly (R2 = 0.48-0.51 for a recommendation), but when crop data were used, they performed significantly better (R2 =

0.75-0.79 for support). Potato output projection is significantly improved over methods that do not make use of cultivar data when ML algorithms are used to combine high-spatial-goal UAV images with cultivar data. By incorporating more precise crop information, details of the soil and landscape, administrative data, and robust machine learning calculations, additional research to improve potato output prediction is likely to be carried out [10].

3. METHODOLOGY

This area of research has numerous issues. Crop conjecture calculation results are as of now satisfactory; however, they could be moved along. A better yield anticipation model that addresses these issues is provided by this study and the proposed study is explained in Fig. 1. The prediction method is based on two main approaches: sequence and feature selection (FS). Before applying FS techniques to balance a collection, it is used to evaluate processes.

- Just data scores with a serious level of worth in concluding the model's final product ought to be incorporated to forestall obvious duplication and further develop ML model accuracy.
- In terms of forecast accuracy, an ensemble method performs better than a prior categorization strategy (Fig. 1).

3.1 KNN

The abbreviation for "K-Nearest Neighbor" is "KNN." It is an AI-overseen calculation. Relapse and issue clustering can both be managed using this strategy. The number of factors with closest neighbors that can be predicted or categorized is represented by the character "K."

3.2 Naive Bayes

Characterization relies on the probability classification of Naive Bayes. Probability models with a lot of flexibility suspicions are used to prove it. Frequently, conceptions of freedom have no bearing on reality. As a result, people think they are stupid.

3.3 Bagging Classifier

A bagging classifier is a meta-assessment that connects individual hypotheses by applying basic classifiers to random segments of the underlying dataset. through the typical or polling form). By incorporating randomization into its growth interaction and then forming a group from it, this kind of meta-assessor is frequently utilized to minimize the difference between itself and a black-box assessor like a decision tree.

3.4 Random Forest

In terms of regulated learning, Random Forest is an essential machine learning technique. It could be applied to ML issues like planning and recurrence. It is based on the idea of accumulation learning, which involves combining several classifications to solve a difficult problem and improve the model's display. A predictor known as Random Forest "works on the extended exactness of that dataset" by "taking the normal of multiple decision trees on diverse subsets of the supplied dataset," as the name of the model suggests. Instead of relying solely on a single chosen tree, the dispersed woods predict the outcome by combining readings from each tree and making use of many projections.

3.5 Decision Tree

Decision trees employ a variety of approaches when deciding whether to divide a hub into at least two sub-hubs. The uniformity of the sub-nodes is enhanced by their appearance. To put it another way, the center becomes cleaner as it gets closer to the target variable.

Support Vector Machine: The well-known Overseen Learning technique known as the Support Vector Machine (SVM) is used for backsliding and portrayal. However, most of its ML applications address classification issues. For n-layered space design, the objective of the SVM method is to determine the ideal line or judgment limit so that subsequent data can be easily added to the correct classification. The simplest and most logical limit is a hyperplane.

Gradient Boosting: In relapse and order applications, gradient boosting, a machine learning technique, is frequently utilized. A collection of unreliable forecast models, most commonly decision trees, is known as an anticipation model. When a decision tree plays the role of the helpless student, gradient-boosted trees are produced. This method does not always outperform random trees. The subsequent development of the model of a gradient-boosted forest is comparable to that of previous aid methods, but it goes above and

beyond by allowing for the development of any visible disaster capacity.

3.6 Voting Classifier

A machine learning predictor known as a voting classifier can predict the outcomes of multiple fundamental models or estimators. Each estimator result's sum parameter could be matched to polling choices.

4. IMPLEMENTATION

The data set is gathered from the Kaggle dataset which is publically available. After collecting the data set, perform the pre-processing of the data to remove noise elements, missing values, and duplicated values. Then building the module which contains the following steps.

- Exploration of data: Data will be entered into the system in this section.
- Information will be received and handled by this program.
- Information will be divided into learn and evaluate segments using this tool.
- The most common way to create a model with or without including selection is referred to as model age. Some of the algorithms used are SVM, Gradient Boosting, Naive Bayes, KNN, Bagging Classifier, Random Forest Decision Tree, and Voting Classifier and a comparative analysis of these algorithms is in Fig. 2. The precision of the algorithm was found.
- Registration and login for users: You must first register before joining to use this feature.
- An estimated output will be produced by using this instrument.
- Prediction: A certain level of expected respect is shown.

4.1 Recursive Feature Elimination

It is a feature selection method used in machine learning to find the dataset's most crucial features for a certain task. The RFE algorithm re-fits the machine learning model using the smaller feature set after successively deleting the dataset's least significant features.

Synthetic Minority Over-sampling Technique: This data augmentation strategy is employed in machine learning to address the problem of class imbalance. Prediction accuracy can be improved while bias is reduced.

Random Over-Sampling Examples. It is a data augmentation technique used in machine learning to address class imbalance problems, like the SMOTE algorithm. Majority Weighted Minority Over-Sampling Technique. It is a data augmentation technique used in machine learning to address class imbalance problems, like the SMOTE and ROSE algorithms.

The classification ensemble exhibit that the model is more effective.

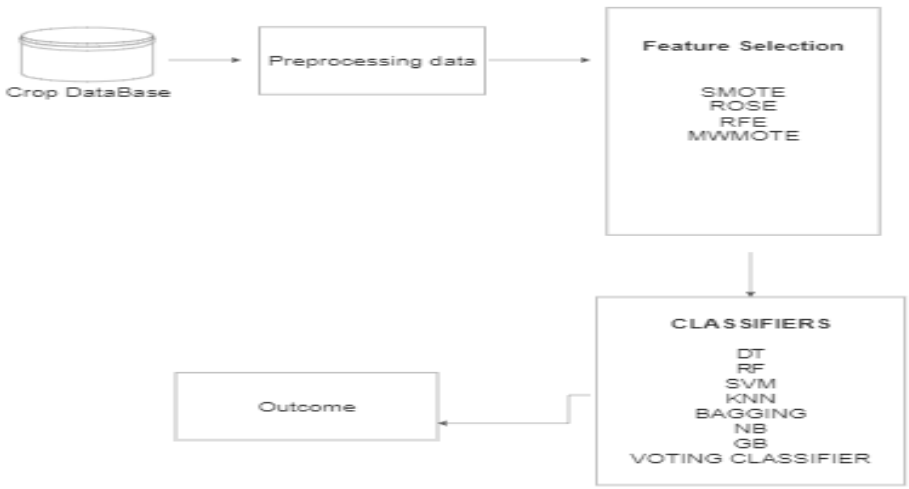

Fig. 1. Framework of the proposed system

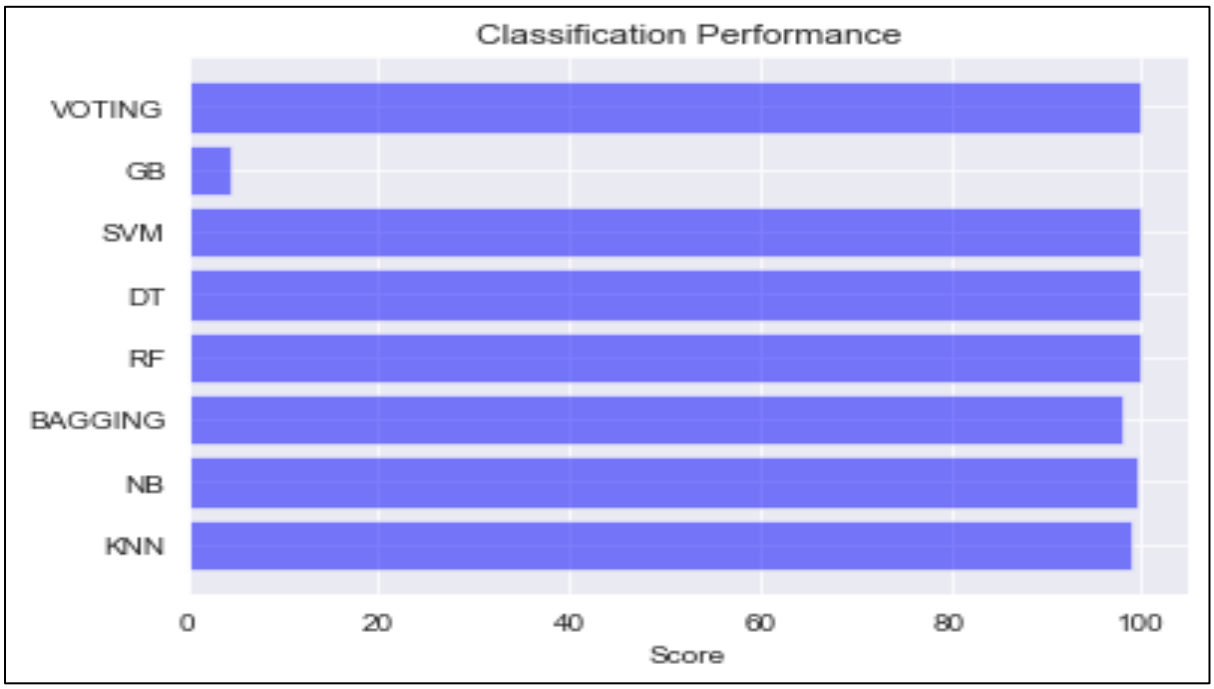

Fig. 2. A comparative analysis of different classification algorithms

5. DISCUSSION AND CONCLUSION

In agribusiness, it is difficult to anticipate crop growth. A variety of element estimation and order methods were utilized in this study to determine the output size of plant growth. From the results, we found out that Random Forest, Support Vector Machine, and Decision Tree give equally good accuracies and, in our paper, we have combined these three algorithms under the voting classifier algorithm to get the best results. The results show that a group approach along with Synthetic Minority Oversampling Technique performs better than the current order methodology in terms of forecast precision. Farmers and nations might find it easier to coordinate their developing endeavors assuming they know where to assemble potatoes, grains, and other energy sources. The use of cutting-edge prediction methods could yield substantial financial rewards.

6. FUTURE ENCHANCEMENTS

As a future enhancement, the current model can be improved by gathering more accurate real-time information regarding crops and various soil and weather parameters with the help of IoT sensors, drones combined with deep learning techniques can enhance the accuracy and precision of the model.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Jahan R. Applying naive Bayes classification technique for classification of improved agricultural land soils, Int. J. Res. Appl. Sci. Eng. Technol. 2018;6(5): 189–193.
2. Sawicka BB, Krochmal-Marczak B. Biotic components influencing the yield and quality of potato tubers. Herbalism. 2017; 1(3):125–136.
3. Sawicka B, Noaema AH, Gáowacka A. The predicting the size of the potato acreage as a raw material for bioethanol production, in alternative energy sources, B. Zdunek, M. Olszáwka, EDS. Lublin, Poland: Wydawnictwo Naukowe TYGIEL. 2016;158–172.
4. Sawicka B, Noaema AH, Hameed TS, Krochmal-Marczak B. Biotic and abiotic factors influencing on the environment and growth of plants, (in Polish), in Proc. Bioróżnorodność Środowiska Znaczenie, Problemy, Wyzwania. Materiały Konferencyjne, Puławy; 2017. Available:https://bookcrossing.pl/ksiazka/321192

5. Myers RH, Montgomery DC, Vining GG, Borror CM, Kowalski SM. Response surface methodology: A retrospective and literature survey. J. Qual. Technol. 2004; 36(1):53–77.
6. Muriithi DK. Application of response surface methodology for optimization of potato tuber yield. Amer. J. Theor. Appl. Statist. 2015;4(4)300–304.
DOI: 10.11648/j.ajtas.20150404.20
7. Marenych M, Verevska O, Kalinichenko A and Dacko M. 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. 2014;16(2):183–188.
8. Olędzki JR. The report on the state of remote sensing in Poland in, (in Polish), Remote Sens. Environ. 2015;53(2):113–174.
9. Grabowska K, Dymerska A, Poáarska K and Grabowski J. Predicting of blue lupine yields based on the selected climate change scenarios. Acta Agroph. 2016; 23(3):363–380.
10. Li D, Miao Y, Gupta SK, Rosen CJ, Yuan F, Wang C, Wang L and Huang Y. Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning. Remote Sens. 2021;13(16):3322.
DOI: 10.3390/rs13163322

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/98549