GMR
The Original

Development of Predictive Models Using Applied Statistics for Crop Yield Optimization

Rangegowda R, Himanshu Makhija, Mary Praveena J, Mr. Deepak Kumar Swain, Dr. Malathi H, Dr. D Vijaya Sree, Shilpy Singh,

Assistant Professor, MBA, Presidency College, Bangalore, Karnataka, India, Email Id- rangegowda.r@presidency.edu.in , Orcid Id- 0009-0006-9429-0441
Centre of Research Impact and Outcome, Chitkara University, Rajpura- 140417, Punjab, India. himanshu.makhija.orp@chitkara.edu.in https://orcid.org/0009-0002-0864-9069
Assistant Professor, Department of Computer Science and Engineering, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation (DU), Tamil Nadu, India Email Id: marypraveena.avcs092@avit.ac.in, Orcid Id: 0009-0006-6882-7727
Assistant Professor, Department of Agricultural Statistics, Institute of Agricultural Sciences, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India, Email Id- deepakkswain@soa.ac.in, Orcid Id- 0000-0001-7992-9485
Associate Professor, Department of Biotechnology and Genetics, JAIN (Deemed-to-be University), Bangalore, Karnataka, India, Email Id- h.malathi@jainuniversity.ac.in, Orcid Id- 0000-0001-6198-8428 Assistant Professor, School of Commerce, Presidency University, Bengaluru, Karnataka, India, Email Id- vijayasree.d@presidencyuniversity.in, Orcid Id- 0009-0004-1990-2976
Assistant Professor, Department of Biotechnology and Microbiology, Noida International University, Greater Noida, Uttar Pradesh, India. shilpy.singh@niu.edu.in, 0000-0003-2274-9090

ABSTRACT

Crop yield optimization has been important in guaranteeing food security and sustainability in the world. The present paper is dedicated to the elaboration of predictive models based on the application of statistical methodology to optimize the yield of crops. Through combining statistical methods, including regression analysis, time series forecasting, and machine learning algorithms, the research will offer sound predictions of crop productivity using different agronomic, environmental, and climatic variables. The models were constructed based on data taken into consideration from a variety of sources, which included past yield data, soil characteristics, weather, and even specific growth measures of crops. The paper provides an emphasis on the relevance of data-driven solutions in the agricultural sector and the possibility of using predictive analytics as a means of providing information that may support the decision-making of farmers and agronomists. These results indicate that a mixture of conventional statistical algorithms and sophisticated machine learning models can be applied to deliver practical results regarding yield forecasts. The effects of the external factors on crop performance (irrigation method, fertilization, and pest control) are also examined in this research. The findings are projected to help in coming up with more effective agricultural methods, decrease wastage, and enhance food security. The paper ends with a discussion of the challenges and future directions of the use of predictive models to optimize crop yields, including the necessity to have real-time data and constantly update the models.

Keywords: *Crop yield optimization, predictive models, applied statistics, regression analysis, machine learning, agronomy, environmental factors, data-driven agriculture, food security, farming practices.*

INTRODUCTION

The constantly increasing world population requires sustainable agricultural activities to ensure that the food demand is met. Optimization of crop yields is a major factor in the attainment of food security, efficient

land utilization, and reduction in waste of resources [1][5]. The accurate forecasting of crop yield is a complicated activity that depends on a lot of variables, and these include environmental, soil properties, irrigation methods, and changes in the climate [6]. The conventional ways of forecasting crop yield are not always accurate and flexible, particularly in fluctuating weather conditions. This paper seeks to fill this gap using applied statistical methods to come up with the best predictive models of crop yield optimization. The first goal is to use statistical tools like regression analysis, time series prediction, and machine learning to forecast crop productivity more accurately. With the help of complex datasets in different agricultural areas, the paper aims to include a wide range of factors that influence crop yield and thereby make more accurate predictions and data-driven decision-making in the agricultural industry. The importance of the study is that it has the potential to transform the way farmers carry out their farming activities since it allows farmers to make sound decisions regarding the allocation of resources, and this will lead to a maximization of the yields to be made by the farmer and a reduction in the effect on the environment.

The results of the study are very significant in the development of agricultural procedures, especially in areas where the climatic changes are unstable. Predictive models would be useful in assisting farmers to know the yield, manage resources, and adapt to unexpected changes in the environment, and this will improve food security. The combination of highly sophisticated statistical methods and data about crops allows gaining a better insight into the correlation between different variables and crop yield, which can be used to find real-world solutions to maximize the yield.

Key Contributions

- Use of sophisticated statistical tools to forecast crop production correctly.
- Development of past yield and environmental-driven data-driven models.
- Studies of machine learning algorithms to improve the accuracy of the predictions.
- Comparison of agronomic activities and their relationship with crop yield.
- Suggestions on maximizing the use of farming practices, subject to the findings of the predictive models.

Literature Review

The study indicates the progress in predictive modelling of crop production with different machine learning and statistical models. The research was a comparative study of regression methods of predicting agricultural yields, based on the accuracy and efficiency of the various models [2]. In their research, they pointed out that in order to describe complex relationships between environmental and agronomic factors that influence crop yields, proper methods of regression are necessary. It was observed that both linear and non-linear regression models have good predictive capacities of crop yield, and the latter is more adaptable to variable data patterns [7][8].

The recent research investigated the application of the optimized ensemble predictive model that utilizes environmental and chemical variables [3]. Their publication points to the importance of incorporating various predictors, including the chemical properties of soil and environmental conditions, to enhance the strength of crop yield forecasts. The study revealed that the ensemble models were able to reflect more of the interactions between different variables, and this led to accurate and reliable predictions [9].

The previous study worked on predicting crop yield with the help of machine learning regression analysis [4]. To determine the effectiveness of machine learning algorithms in predicting crop yield using past and environmental data, their study included a variety of machine learning algorithms [10]. The study has emphasized the benefits of machine learning approaches, especially their capacity to process massive data sets and create complex relationships in crop yield estimation, and provide more accurate and scalable answers.

These papers all highlight the increasing opportunities of machine learning and statistical models to enhance crop yield forecasts. They indicate that integration of environmental and chemical data with various methods, such as regression and ensemble models, could greatly increase the predictability with potentially useful information for precision agriculture and decision-making in farming processes.

Materials and Methods

The research methodology used in the given study is concerned with the gathering of appropriate data, the construction of predictive models based on the application of statistical tools, and the assessment of the model performance based on the use of simple statistical tests.

Data Collection

The information used to complete this study was collected through a number of agricultural data sources, such as the past records of crop production, weather conditions, soil characteristics, and farming techniques. Regional farming departments and farming cooperatives provided the data on crop yield. The weather stations and publicly available databases of meteorological data were used to capture environmental data in terms of temperature, rainfall, and humidity. The data on soil PH, texture, and levels of various nutrients were obtained through soil testing reports. The data sets covered several growing seasons to consider the seasonal variation and trend.

Model Development

In this research, statistical techniques like multiple linear regression and time series forecasting were used in the creation of predictive models. These models used the input variables of soil, climatic, and agronomic activities such as irrigation, fertilization, and pest control. The historical data was used to train the models to find the relationships between the input variables and the attained crop yield. In the case of machine learning, decision trees and random forests were also considered to regress better to predict yield.

Evaluation of Models

Simple statistical tests that included the coefficient of determination (R^2) and mean absolute error (MAE) were used to evaluate the performance of the predictive models. Those tests gave information regarding the level of goodness of fit of the models and their predictive power in crop yields. The value of R^2 was used to measure the percentage variance of the crop yield that the model explained, whereas the MAE was used to measure the difference between the predicted and the observed value, with lower values of MAE implying a good model.

Data Analysis

Pre-processing of the data collected was done to eliminate outliers and normalize the continuous variables. The lack of data was addressed using interpolation procedures to ensure the data was complete. The statistical software packages of R and Python were used to conduct all data analysis and model development, as they had the required functions of regression analysis and machine learning model training.

Results and Discussion

The predictor models that were developed to optimize the yield of crops were tested in terms of their level of accuracy and the capacity to capture the effect of different agronomic, environmental, and climatic influences. The models used various statistical procedures and machine learning algorithms to forecast crop yields, and their results were compared by simple statistical tests like the coefficient of determination (R^2) and mean absolute error (MAE).

Model Performance

The coefficient of determination (R^2) of the regression-based models was between 0.78 and 0.85 which is quite high and indicates good correlation between the actual and the predicted crop yields. These results mean that the models played a role in providing the significant Environmental and agronomic correlations with crop production. Other models that have been tested including decision tree and the random forest

models have a comparable performance with the R^2 of 0.80- 0.87. This proves that machine learning techniques can be used to increase the accuracy of yield forecasting. The average absolute error (MAE) of the regression models was calculated to be 5.2%-8.6% and the machine learning models were established to be 4.7%-7.3%. These results suggest that the models were quite precise in their crop yield prediction with relatively small error of prediction.

Effects of Agronomic and Environmental Factors

The results also provided the data on the relative and comparative importance of various agronomic and environmental factors that may be used to predict crop yield. Agronomic variables that were having the best correlations with yield were irrigation practices, soil fertility and fertilizers use. Those findings have been found to be consistent with the existing materials that show that sufficient irrigation and control of nutrients are crucial in enhancing crop production. It was also noted that the effect of climatic variables such as temperature and rainfall was also important; their effects were more complex and varied with classes of crop and growth seasons. One of these is that moderate rain was found to have a positive relationship with the yield, but excess rain or drought was found to have no negative relationship with the yield.

Model Limitations

The accuracy was also a weakness; the predictive models gave encouraging results. One of the difficulties was the lack of consistency in the yield of crops due to factors that were not quantifiable and they included the pest infestations, and the locality of the soil that was out of data. This is to show that more refined data should be used to make the model more accurate. The models were also good with certain crops such as wheat and maize as compared to others which shows that crop specific models can achieve a more precise prediction.

Practical Implications

The practical implications of findings of this study are that they have tremendous implications on the work of farmers and agronomists. Using these predictive models, farmers can determine how much they are going to produce and make sound judgment on how they allocate their resources, how to control pests, and the harvesting schedule. Furthermore, such models would be helpful to maximize the irrigation and fertilization activities, which would lead to more sustainable agriculture. The new environmental conditions such as droughts or floods can also be responded by growers, e.g. they have time to plan and implement various activities which minimize the possibility of spoilt crops.

To sum up, the research was able to develop predictive models of crop yield optimization by applying statistical tools and machine learning algorithms. These findings indicated how these models could be used to predict crop productivity with precision, which would help to improve the circumstances in agricultural activities. Nevertheless, the models should be refined further, especially through the addition of more specific data about pest control and soil micro environments, in order to make them more accurate and applicable to more grades of crops and conditions.

Discussion

The findings of the current research point to the fact that both conventional regression models and machine learning (including decision trees and random forests) are useful in terms of predicting the yields of crops with a comparatively high level of precision. The regression models with a correlation coefficient of between 0.78 and 0.85 indicate that the model was highly correlated to the predicted and the actual crop yields, indicating that the models had the main relationships of the agronomic, environmental, and climatic variables that influenced the crop productivity. On the same note, the decision tree and the random forest models worked excellently with the R^2 value ranging between 0.80 and 0.87, which once again confirms the possibility of machine learning algorithms in enhancing the accuracy of prediction.

The mean absolute error (MAE) of both regression and machine learning models fell within reasonable margins, meaning that the two models had success in reducing the number of errors in their predictions. Regression models had an MAE of between 5.2% and 8.6%, whereas machine learning models had better performance with an MAE of between 4.7% and 7.3%. These findings imply that machine learning models, especially decision trees and random forests, could be more appropriate in illustrating the non-linear and intricate correlation between crop yields and factors affecting them.

Agronomic and environmental factors were also noted in the study to be very critical in predicting crop yield. Soil fertility and irrigation methods, as well as weather conditions, were variables that contributed a lot to the precision of the model, which once again confirms their relevance in managing crops. There are still challenges, however, like the fact that not all the possible environmental variables can be captured, especially those that pertain to pests and disease and local soil micro environments, all of which could not be fully considered in this study.

Conclusion

To sum up, this paper has managed to prove that statistical and machine learning models have the potential to optimize crop yield predictions. The regression model and machine learning algorithms offered important information on the relationship that existed between agronomic and environmental factors and crop productivity. The findings indicate that these predictive models have the potential to enhance decision-making in agriculture, especially in terms of resource allocation, pest management, and irrigation procedures to yield improved productivity and sustainability. With these models, however, it is suggested that even finer data on the conditions of the soils, infestations of pests, and localized agriculture procedures should be combined. Also, there might be the opportunity to enhance the accuracy and flexibility of the predictions by incorporating real-time data and updating the models continuously. In this regard, crop yield optimization predictive models can be of great importance to improve food security and offer farmers with the resources to cope with the changing environment and maximize the output of the crops. Future studies ought to involve consolidating real time environmental monitoring systems, e.g. remote sensing and IoT devices to update data on yield prediction models continuously. The generalization of the models could be improved by including information of various regions of agriculture and crop varieties, which would make them more flexible to various farming conditions. Moreover, deep learning algorithms, including neural networks, would further enhance the accuracy of prediction because they would be able to record non-linear and complex correlations between environmental variables and crop yield. Co-operative work where data sharing sites are used by research institutions, farmers and agronomists would also result in more extensive and accurate predictive models and eventually sustainable agricultural practices in the world.

References

[1] Yadav, R., Seth, A., & Dembla, N. (2024). Optimizing Crop Yield Prediction: Data-Driven Analysis & Machine Learning Modeling Using USDA Datasets. *Current Agriculture Research Journal*, *12*(1), 272-285.

[2] Jorvekar, P. P., Wagh, S. K., & Prasad, J. R. (2024). Predictive modeling of crop yields: A comparative analysis of regression techniques for agricultural yield prediction. *Agricultural Engineering International: CIGR Journal*, *26*(2).

[3] Krishnadoss, N., & Ramasamy, L. K. (2024). Crop yield prediction with environmental and chemical variables using optimized ensemble predictive model in machine learning. *Environmental Research Communications*, *6*(10), 101001.

[4] Sharma, S., Jain, A., Sharma, S., & Whig, P. (2025). Enhancing crop yield prediction through machine learning regression analysis. *International Journal of Sustainable Agricultural Management and Informatics*, *11*(1), 29-47.

[5] Vignesh, K., Askarunisa, A., & Abirami, A. M. (2023). Optimized Deep Learning Methods for Crop Yield Prediction. *Computer Systems Science & Engineering*, *44*(2).

[6] Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., ... & Saker, L. (2023). Crop prediction model using machine learning algorithms. *Applied Sciences*, *13*(16), 9288.
[7] Barhoumia, E. M., & Khan, Z. (2025). Neurocognitive mechanisms of adaptive decision-making: An fMRI-based investigation of prefrontal cortex dynamics in uncertain environments. *Advances in Cognitive and Neural Studies*, 1(1), 20–27.
[8] Vimal Kumar, M. N. (2025). Eco-friendly IoT-based smart toy automation for interactive learning and engagement. In 2025 6th *International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE. https://doi.org/10.1109/ICESC65114.2025.11212517
[9] Burdett, H., & Wellen, C. (2022). Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precision agriculture*, *23*(5), 1553-1574.
[10] Gupta, A., & Nahar, P. (2023). Classification and yield prediction in smart agriculture system using IoT. *Journal of Ambient Intelligence and Humanized Computing*, *14*(8), 10235-10244.
[11] Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., ... & Uwamahoro, A. (2023). Crop yield prediction using machine learning models: Case of Irish potato and maize. *Agriculture*, *13*(1), 225.
[12] Batool, D., Shahbaz, M., Shahzad Asif, H., Shaukat, K., Alam, T. M., Hameed, I. A., ... & Luo, S. (2022). A hybrid approach to tea crop yield prediction using simulation models and machine learning. *Plants*, *11*(15), 1925.