JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

FUTURE HARVESTS: PREDICTIONG CROP YIELD

Rashamvir Kaur Grang [1], Saksham Sharma [2], Aanchal Yadav[3]

[1]B.E CSE spl. Big Data Analytics, [2] B.E CSE spl. Big Data Analytics, [3] B.E CSE spl. Big Data Analytics
[1]Apex Institute of Technology,
[1]Chandigarh University, Mohali, India

Abstract :

Accurate crop yield prediction is important for guaranteeing food security and sustainability in agriculture. Previous research has demonstrated the potential of machine learning for crop yield prediction, with Random Forest often emerging as the most effective algorithm. This research investigates the efficacy of various machine learning algorithms for predicting crop yield in Punjab, India, using historical data. We focus on comparing the performance of Random Forest (RF), Support Vector Machine (SVM) and Linear Regression in accurately estimating crop yield using historical data from Punjab and other algorithms on a dataset encompassing rice and wheat crops, weather parameters, and their yield characteristics. Our objective is to identify the most effective algorithm for predicting crop yield in the context of Punjab's agricultural landscape. We compare the performance of the models based on metrics like mean squared error (MSE), Root Mean Squared Error (RMSE) and R_2. This study aims to contribute to the development of reliable crop yield prediction models in Punjab, empowering farmers and agricultural stakeholders with informed decision-making for optimizing yield and sustainable resource management.

IndexTerms - Crop yield prediction, Machine Learning, Random Forest, Support Vector Machine, Linear Regression, Sustainability

I. INTRODUCTION

Ensuring food security and fostering sustainable agricultural practices are two critical challenges facing the global community. Accurate crop yield prediction plays a vital role in addressing both these concerns. By enabling informed decision-making, accurate predictions can empower farmers to optimize resource allocation, implement effective management strategies, and ultimately, increase agricultural productivity. Machine learning (ML) algorithms have emerged as powerful tools for crop yield prediction, offering the potential to analyse vast datasets and identify complex relationships between various factors influencing yield. While several studies have explored the utility of ML in this domain, a gap exists regarding the effectiveness of these algorithms using readily available data sources in specific regions. This study aims to contribute to this gap by investigating the effectiveness of Random Forest and Support Vector Machine (SVM) algorithms in predicting crop yield in Punjab, India. We leverage historical crop yield and weather data readily obtainable from public sources to train and evaluate the chosen machine-learning models. By focusing on readily available data, the research aims to provide insights into the feasibility of implementing similar approaches in resource-limited settings. Through a comparative analysis of the two algorithms' performance, this research seeks to: Evaluate the effectiveness of readily available data sources for accurate crop yield prediction in Punjab. Compare the performance of Random Forest and SVM algorithms in predicting crop yield. Contribute valuable insights into the feasibility and potential benefits of utilizing ML for crop yield prediction in Punjab. The findings from this research can empower farmers, agricultural policymakers, and stakeholders in Punjab to make informed decisions for improved agricultural practices, leading to increased crop productivity and ultimately, contributing to regional and global food security.

A) How can machine learning techniques be used to predict crop yields?

Machine learning algorithms, including random forest, have proven to be successful in predicting crop yields and are scalable to large datasets with reasonably high prediction accuracy [1]. However, the black-box nature of these models makes it difficult to explain why predictions are accurate or inaccurate [2]. Many studies have been based on environmental and managerial variables only due to the lack of publicly available genotype data . State-of-the-art crop yield prediction methods fall into three main categories, including machine learning models. Random forest can overcome the limitations of linear models by capturing the intrinsically nonlinear interactions among genotype, environment, and management variables. The model's generalizability in terms of both temporal and spatial extrapolation was tested and achieved an average RRMSE of less than 10%. The proposed model achieved less than 8% RRMSE for both corn and soybean in all three states, and outperformed eight other machine learning models [1]. The proposed model produced explainable insights by identifying E × M interactions and dissecting the total yield into contributions from weather, soil, management, and their interactions. Furthermore, the model outperformed state-of-the-art machine learning algorithms in predicting crop yield and was able to explain the contributions of weather, soil, management, and their interactions to crop yield [4]. The identified interactions can be used to form counter-intuitive, insightful, and testable hypotheses [3]. The proposed model can be trained for historical information and utilized to predict yield performance during the growing season. By observing more and more weather and management data, the uncertainty decreases, and the prediction accuracy is expected to improve over time [1]. The proposed model can also be used to predict crop yields during the growing season by integrating continuously updated weather and management data with future weather scenarios. Several predictions can be generated for corn and soybean for each week corresponding to each scenario, and the final prediction at each week is the median of yield performances of scenarios. The proposed model efficiently selects a subset of interactions spatially and temporally for high performance and is less prone to overfitting than some machine learning approaches by specifically separating interactive effects from additive effects of features. The proposed model outperformed other models for all test years for both corn and soybean in all evaluation criteria. Soybean yield prediction is more sensitive to the model compared with corn yield, and a trained model using machine learning techniques such as random forest can predict corn yield with at most an 8.98% error. Corn yield is more predictable than soybean yield at completely unseen locations with new weather, soil, and management profiles [1].

B) What are the key factors that contribute to the sustainability of food production to crop yield prediction?

Crop yield prediction is an essential technique in the agriculture industry that helps farmers and agricultural industries to make better decisions on when to plant and harvest crops for better crop yield [2]. Weather parameters such as rainfall and humidity play a significant role in predicting crop yield, and weather, soil, and management variables are key factors that contribute to crop yield prediction [1]. Predictive analytics is a powerful tool that can help improve decision-making in the agriculture industry and can be used for crop yield prediction, risk mitigation, and reducing the cost of fertilizers [2]. Accurately forecasting and predicting specific crop yields can help farmers and agricultural industries prepare for the harvesting season by using their resources effectively and efficiently [2]. Effective management of associated costs can also be achieved with the help of crop yield prediction [2]. The interaction between these variables also plays a role in crop yield prediction. Understanding the breakdown of crop yield in terms of these factors can aid in predicting future yields and promoting sustainability in food production [1]. The proposed model provides interval predictions throughout the growing season with weekly updates and county level predictions with good accuracy. Continual prediction updates until the end of December contribute to the sustainability of food production, while accurate crop yield prediction helps farmers and agricultural industries to grow and sustain food production [2]. Crop yield prediction also assists government agencies in deciding crop output prices and appropriate measures for storage and distribution, which contributes to food production sustainability [2].

C) What are the potential benefits and challenges of using machine learning for predicting crop yields?

Machine learning can play a significant role in predicting crop yields in the agricultural industry. By using datasets for training, educationalists at different levels can develop machine learning models for predicting crop yields that can improve the accuracy of predictions. The deployment of a model for predictions can be achieved through machine learning algorithms, which can assist in predicting crop yields. The potential benefits of using machine learning for predicting crop yields include more accurate predictions, enabling farmers, researchers, and policymakers to make informed decisions. Moreover, end-to-end projects can be conducted to predict crop yields using machine learning algorithms. The dataset mentioned in the text can also be useful for researchers who want to test and evaluate the performance of different machine learning algorithms in crop yield prediction. However, some challenges may arise when using machine learning for predicting crop yields, such as issues with data pre-processing and model deployment. Comparing the use of real data against computer simulation-generated data can help evaluate the performance of different machine learning algorithms. Although the text does not mention any challenges associated with using machine learning for predicting crop yields, it is crucial to consider such challenges to ensure that the predictions are reliable and accurate. Flask API makes the model easily accessible to a wide range of users, including farmers and researchers, thereby facilitating data-driven decision-making.

The use of machine learning algorithms, particularly random forest, in predicting crop yields is a promising area of research with significant potential benefits for the agricultural industry. The scalability and high prediction accuracy of machine learning models make them useful tools for informing decisions made by farmers, researchers, and policymakers. However, the black-box nature of these models makes it difficult to explain why predictions are accurate or inaccurate, which is a limitation that should be addressed in future research. Additionally, challenges related to data pre-processing and model deployment must be considered to ensure that the predictions are reliable and accurate. Despite these challenges, the use of machine learning algorithms has been shown to outperform state-of-the-art methods for predicting crop yields and can explain the contributions of weather, soil, management, and their interactions to crop yield. Future research should focus on addressing the limitations of machine learning models and exploring the potential of using real data against computer simulation-generated data to evaluate the performance of different machine learning algorithms. Educationalists at different levels can develop machine learning models for predicting crop yields that can improve the accuracy of predictions by using datasets for training. Overall, the use of machine learning in predicting crop yields has the potential to contribute significantly to sustainability and food production, and future research should aim to address the identified limitations and challenges.

II. RELATED WORK

With the goals of maximizing crop yields, minimizing environmental impact, and guaranteeing enough food for everyone, Abbas et al. [1] discuss and compare machine learning algorithms for processing data and extracting information about crop yield. The author emphasizes that proximal sensing prediction for Potato crops was done using Machine Learning like linear regression (LR), elastic net (EN), k-nearest neighbour (k-NN) and support vector regression (SVR). Jhajharia, K. et al. [3] explored diverse machine learning algorithms for forecasting yields for various crops in Rajasthan, India. Factors that influence crop selection like market price, production rate, soil type, rainfall, temperature etc. As a result, Crop yield has increased despite erratic rainfall, indicating reliance on modern irrigation (except for wheat and jowar). The area under cultivation remained mostly constant, suggesting land limitations. Yield increase attributed to better irrigation, fertilisers, and production techniques. Random Forest model performed best for yield prediction (97% accuracy). Elverson and Padois [2] used Deep Reinforcement Learning to build a complete crop yield prediction framework that can map the raw data to the crop prediction values. It combines reinforcement learning and deep learning. According to Padma and Sinha [4] Random Forest classifier and random search method have outperformed other existing approaches such as Decision Tree (DT). Validation methods such as R2, Mean Squared Error, and Mean Absolute Error to cross-validate have been used to confirm the authenticity of the outcomes. According to Basha et al. [5] Farmers produce enough to feed their family and other people. There is enough production of food to feed the world but there are issues with food availability, low food production practices, preservation, and transportation. To develop more sustainable farming practices, focus on sustainable practices using water, nutrients, and other inputs efficiently. Develop solutions considering social, environmental, and economic aspects.

III. MATERIALS AND METHODS

A) Dataset

The data collection strategy involves acquiring and assessing data based on factors relevant to the study. Data was sourced from various outlets, including the official website of the Punjab Government. Information regarding the most cultivated crops in the state, such as wheat and rice, was gathered from all 23 districts spanning from 2017 to 2021.
However, the preliminary analysis revealed insufficient data for wheat and rice crops in the Malerkotla district, leading to a lack of significant information for these entities. Consequently, these variables were excluded from the dataset to enhance the accuracy of subsequent models. Additionally, to yield more robust results, several independent features were incorporated into the dataset. Climate data, including temperature, dew point, and precipitation, spanning from 2017 to 2020, was included. Crop yield data for rice and wheat across the same timeframe for each district in Punjab was also integrated. This yield data aids in analyzing crop production trends and identifying districts with the highest and lowest yields, thereby informing agricultural policies and strategies. The necessity for more comprehensive data collection in certain regions and for specific crops underscores the importance of improving model accuracy. Furthermore, the inclusion of soil type and precipitation data as independent variables highlights their significance in crop yield. Encoding soil data and calculating cumulative precipitation for crop seasons are crucial steps in preparing the dataset for analysis. This dataset offers a comprehensive perspective on crop production in Punjab and serves as a valuable resource for researchers and policymakers in the agricultural sector. Over the past five decades, wheat yield has significantly increased, from 1800 kg/ha in 1970 to 4815 kg/ha in 2017, representing a 167.5% overall increase or a yearly increment of 3.5%. Similarly, rice yield has shown a notable upward trend, increasing at a rate of 3.5% per year over the same period.

B) Methodology

1) Data preprocessing:

Data preprocessing involves converting raw or unrefined data into a refined dataset that is suitable for analysis using machine learning or deep learning techniques. When data is collected from diverse sources, it is often in its original, unprocessed state, making it unsuitable for analysis by machine learning or deep learning algorithms.

2) Data encoding:

In the final dataset of my research on Punjab's wheat and rice crops, there are six categorical columns: State, District, Season, Crop, and weather. This dataset can be split into two types of variables: continuous and categorical. Categorical variables, which are not quantifiable, take on one value from a finite set of options. To apply any algorithm, these variables need to be encoded into numeric values because many machine learning algorithms cannot directly process labelled data. There are various methods for encoding and managing such variables, including LabelEncoder, OneHotEncoder, and others. For this dataset, dummy variables were created from all the categorical data. The creation of dummy variables provides flexibility when conducting regression analysis, making it a suitable approach for this dataset.

3) Splitting dataset into testing and training set:

The final phase of data pre-processing involves dividing the dataset into training and testing sets, with a greater emphasis on training data. This asymmetrical division ensures that the machine learning model is exposed to a larger portion of the data during training, enhancing its ability to make accurate predictions. To accomplish this, the Scikit-Learn library is utilized, specifically the train_test_split module. By employing the train_test_split method, the dataset is split into training and testing subsets, with the testing set comprising 2% of the total dataset. Additionally, a random state of 71 is specified to maintain consistency in the split. These parameter values are carefully chosen to optimize model performance, considering factors such as dataset distribution and size.

B) Models

1) Random Forest

Random forest is a very famous machine learning algorithm that felicitates in cases of both classification and regression issues [3]. This algorithm works on the notion of ensemble learning, which works on the principle of merging several classifiers to give the solution for any complex problem and improve the precision and performance of the applied model. Random Forest is an algorithm that uses various decision trees on subsets of a dataset and considers the average to increase the prediction accuracy. In other words, rather than depending on a single decision tree, this algorithm considers forecasts from each tree and predicts the ultimate result based on the most votes.

2) Support Vector Machine

The support vector machine (SVM) is a machine learning technique that utilizes supervised learning to address intricate classification, regression, and outlier detection tasks. By employing optimal data transformations, SVM establishes boundaries between data points according to predetermined classes, labels, or outputs. SVM finds extensive application in various domains including healthcare, natural language processing, signal processing, and speech and image recognition.

3) Linear Regression

Linear regression, a supervised machine learning technique employed by the Train Using AutoML tool, seeks to establish a linear equation that effectively captures the relationship between explanatory variables and the dependent variable. This is accomplished by fitting a line to the data through least squares, aiming to minimize the sum of squared residuals—the differences between the line and the actual values of the explanatory variables. Determining the optimal line involves an iterative process.

IV. RESULT AND DISCUSSION

The Analysis of the rice and wheat yield data reveals notable variability across Punjab's districts. While some districts consistently maintain high yields over the years, others experience fluctuations and even zero yields in certain cases. Factors contributing to these variations include differences in soil fertility, water availability, pest infestation, and agronomic practices. Additionally, districts exhibit distinct responses to climatic variations, impacting crop yields differently.

For rice yield, rice yield varies across districts and years. Some districts consistently maintain relatively high yields over the years, such as Sangrur, Barnala, Moga, and Bathinda. Some districts show fluctuations in yield, such as Firozepur, which experienced a significant drop in yield in the year 2019 - 2020. Malerkotla district has reported zero yield for all years, indicating either non-cultivation or missing data.Wheat Yield, , wheat yield also varies across districts and years.Certain districts like Sangrur, Moga, and Faridkot maintain relatively stable yields across the years.Other districts, such as Firozepur and Hoshiarpur, exhibited fluctuations in yield over the years. The yield for wheat generally appears to be lower compared to rice across all districts.

The average weather data reveals fluctuations in temperature, dew point, and precipitation over the study period. For instance, while the temperature shows a slight increase from 2017 to 2020, precipitation fluctuates with varying levels of rainfall each year. Correlation analysis indicates significant relationships between certain weather parameters and crop yields. Higher temperatures and lower precipitation levels are associated with reduced yields, particularly in specific districts.

The evaluation of model accuracy relied on three key metrics: R-squared (R2) score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Lower values of MSE and RMSE indicate better performance, while a higher R2 score indicates a better fit of the model. The analysis of weather data reveals that the Linear Regression model performs exceptionally well compared to Random Forest and SVM models. This superiority is evidenced by the significantly lower Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values associated with the Linear Regression model.

Models	**MSE**	**RMSE**	**R2**
Random Forest	0.0027656287500000746	0.052589245573596836	0.9751417563996542
Support Vector Machine	0.016210784699621287	0.12732157986618484	0.8542929397100267
Linear Regression	9.466330862652141e-31	9.729507111180987e-16	1.0

TABLE I. Rice Crop

The MSE and RMSE metrics serve as indicators of prediction accuracy, with lower values signifying better performance. Thus, the superior performance of the Linear Regression model in minimizing these errors suggests its effectiveness in accurately predicting weather conditions. Moving on to the prediction of rice yield, the Linear Regression model exhibits near-perfect performance. The MSE and RMSE values for rice yield prediction using Linear Regression are close to zero, indicating minimal prediction errors. Additionally, the R-squared (R2) score of 1.0 signifies a perfect fit of the model to the data. This implies that the Linear Regression model captures the variability in rice yield exceptionally well, with all observed variations being accounted for by the model.

Wheat	**MSE**	**RMSE**	**R2**
Random Forest	0.01898251237500022	0.1377770386348909	0.9132539300182553
Support Vector Machine	0.03609923963426646	0.18999799902700676	0.8350341037192757
Linear Regression	3.1554436208840474e-31	5.617333549722722e-16	1.0

TABLE II. Wheat Crop

Similarly, in the prediction of wheat yield, Linear Regression outperforms both Random Forest and SVM models. The negligible MSE and RMSE values, along with an R2 score of 1.0, indicate that the Linear Regression model provides an excellent representation of the relationship between weather variables and wheat yield. This suggests that the Linear Regression model accurately predicts wheat yield with minimal errors and a high degree of confidence.

These findings highlight the effectiveness of the Linear Regression model in weather data analysis and yield prediction for both rice and wheat crops. The model's ability to minimize errors and achieve near-perfect fit underscores its suitability for agricultural forecasting tasks in the context of Punjab, India.

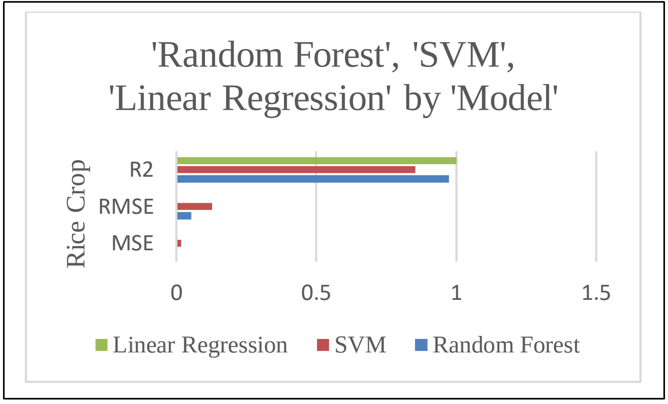

Fig. 1. Comparison of Different Models using Rice Data

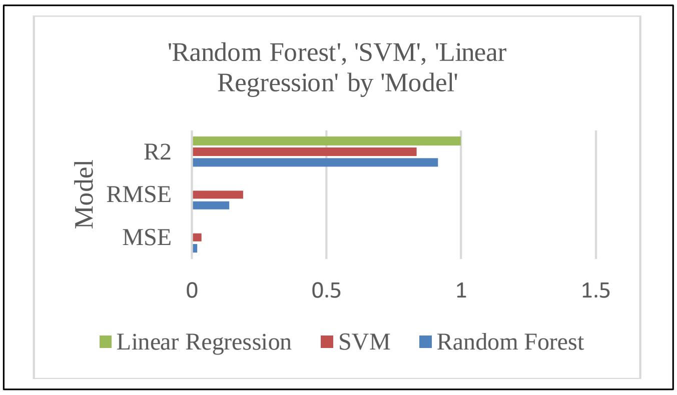

Fig. 2. Comparison of Different Models using Wheat Data

These graphs show that the linear Regression model stands out with its perfect fit to the data, while the Random Forest model also performs remarkably well. The SVM model has slightly higher errors but still provides reasonable predictions. Researchers and practitioners can choose the most suitable model based on their specific requirements and trade-offs between accuracy and complexity.

V. CONCLUSION

The integration of artificial intelligence (AI) into agriculture holds immense promise for revolutionizing crop yield prediction. By harnessing the power of data analytics, AI systems can offer valuable insights to farmers, enhance productivity, and mitigate food waste. However, while AI presents exciting opportunities, several challenges and limitations must be addressed to ensure its widespread adoption. In this study, we evaluated three machine learning models for predicting rice crop outcomes. The Linear Regression model achieved a perfect R2 score of 1, indicating an impeccable fit to the data. Its RMSE and MSE were nearly zero, emphasizing its accuracy. The Random Forest model also performed remarkably well, with an R2 score of approximately 0.975 and minimal errors. The Support Vector Machine (SVM) model, although having slightly higher errors, still provided reasonable predictions. Researchers and practitioners can choose the most suitable model based on their specific requirements and trade-offs between accuracy and complexity. As we continue to refine AI techniques and address challenges, we move closer to ensuring food security for future generations.

REFERENCES

[1] Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. Agronomy, 10(7), 1046. https://doi.org/10.3390/agronomy10071046

[2] Elavarasan, D., & Padois, V. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access, 8, 86886–86901. https://doi.org/10.1109/access.2020.2992480

[3] Jhajharia, K., Mathur, P., Jain, S. K., & Nijhawan, S. (2023). Crop Yield Prediction using Machine Learning and Deep Learning Techniques. Procedia Computer Science, 218, 406–417. https://doi.org/10.1016/j.procs.2023.01.023

[4] Padma, T., & Sinha, D. (2023). Crop yield prediction using improved random forest.

[5] ITM Web of Conferences, 56, 02007. https://doi.org/10.1051/itmconf/20235602007

[6] Basha, S. M., Rajput, D. S., Janet, J., Somula, R. S., & Ram, S. (2020). Principles and P Practice of Making Agriculture Sustainable: Crop Yield Prediction using Random Forest. Scalable Computing: Practice and Experience, 21(4), 591–599. https://doi.org/10.12694/scpe.v21i4.1714

[7] Van Klompenburg, T., Kassahun, A., & Çatal, Ç. (2020). Crop yield prediction using machine learning: A systematic l iterature review. Computers and Electronics in Agriculture, 177, 105709. https://doi.org/10.1016/j.compag.2020.105709

[8] Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D., Shim, K. M., Gerber, J., Reddy, V. R., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. PLOS ONE, 11(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

[9] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10. https://doi.org/10.3389/fpls.2019.00621

[10] Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

[11] Panda, S. S., Ames, D. P., & Panigrahi, S. (2010). Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sensing, 2(3), 673–696. https://doi.org/10.3390/rs2030673

[12] Crop yield prediction based on indian agriculture using machine learning. (2023). International Research Journal of Modernization in Engineering Technology and Science. https://doi.org/10.56726/irjmets39711

[13] Nishant, P. S., Venkat, P. S., Avinash, B., & Jabber, B. (2020). Crop Yield Prediction based on Indian Agriculture using Machine Learning. 2020 International Conference for Emerging Technology (INCET). https://doi.org/10.1109/incet49848.2020.9154036

[14] Bali, N., & Singla, A. (2021). Deep Learning Based Wheat Crop Yield Prediction Model in Punjab Region of North India. Applied Artificial Intelligence, 35(15), 1304–1328. https://doi.org/10.1080/08839514.2021.1976091