Applied Biological Research 26(2): 199-209; (2024)
DOI: 10.48165/abr.2024.26.01.23

PREDICTIVE MODELING FOR SUGARCANE PRODUCTION: A COMPREHENSIVE COMPARISON OF ARIMA AND MACHINE LEARNING ALGORITHMS

Vishwajeet Singh[1], Med Ram Verma*[2] and Subhash Kumar Yadav*[3]

[1]*Directorate of Online Education, Manipal Academy of Higher Education, Manipal – 576 104, Karnataka (India)*
[2]*Division of Design of Experiments, ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi – 110 012 (India)*
[3]*Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow - 226 025, Uttar Pradesh (India)*
*****e-mail: *drmrverma@gmail.com; drskystats@gmail.com*

(Received 22 December, 2023; accepted 14 April, 2024)

ABSTRACT

Accurate prediction of sugarcane yield is essential for trade, economic planning, and sustainable agriculture in India. This study addressed the challenge of forecasting sugarcane yield by evaluating the effectiveness of time series modelling and machine learning algorithms. Leveraging data spanning from 2001 to 2020, the research focuses on predicting the sugarcane yield for the subsequent years. The problem statement revolves around the need for precise yield predictions to inform decision-making in the agricultural sector. Methods employed included the utilization of Autoregressive Integrated Moving Average (ARIMA) for time series analysis and machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machine (GBM). The analysis encompassed sugarcane yield data spanning multiple years, with predictions extending for a specified duration. Through analysis of temporal patterns and dependencies within the sugarcane yield time series data using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), the study optimized the predictive models. Results indicated that ARIMA outperformed machine learning algorithms, exhibiting superior performance with a root mean square error of 36700.68 and a minimum AIC value of 456.7. The study emphasizes the significance of accurate yield predictions for agricultural planning and decision-making, highlighting the implications for sustainable crop management and the fortification of Indian sugar industry. The study affirms the importance of informed decisions facilitated by accurate yield predictions in resilient agricultural sector. Overall, this study contributes to the advancement of sugarcane yield prediction, offers practical insights for stakeholders and policymakers in India's agricultural landscape.

Keywords: ARIMA, gradient boosting machine, machine learning methods, random forest, sugarcane yield prediction, support vector machine

INTRODUCTION

Sugarcane has substantial impact on India's trade, economy and lifestyle. It is the most extensively cultivated cash crop in India and annually contributes about ₹ 27,000 crores to the country's economy. Sugar production, India's second-largest agriculture-based business after cotton production, is totally dependent on sugarcane harvest. More than 5 crore farmers and labourers are employed in India's sugar and sugarcane industries alone. In India, where sugarcane is planted on over 5 million ha, sugar production ranks 2nd in the world (Sneha and Bhavana, 2023). An estimated average 35.5 million tons

(m t) of sugarcane are harvested annually in India, yielding 30 m t sugar (Bhatt *et al.*, 2023). There are about 732 sugar mills in India, which contribute about 5 KW of renewable energy to the national grid (Bhatt and Singh, 2021). The Government of India is investing heavily in ethanol blending to become self-sufficient in fuel business, which is solely dependent on the output of sugarcane (Final Report of the Task Force on Sugarcane and Sugar Industry, NITI Ayog, India, 2020). Sugarcane is an important alcohol source for biofuel manufacture. Presently, there is increased interest in using sugarcane for bio-based goods such as bio-plastics and bio-chemicals to aid nation's shift to environmentally acceptable alternatives. Sugarcane's economic importance as a key cash crop supporting millions of farmers and employees has also drawn attention to the industry, influencing legislative decisions and market dynamics (Mishra *et al.*, 2021). As a result, effective systems that provide fast and accurate information on sugarcane production and growth conditions at regional and global sizes are in high demand. The management of import-export, marketing, and timely production decisions, as well as distribution, pricing, and other critical agricultural strategies, depend on accurate crop production forecasts. Although the current crop yield prediction models work reasonably well, there is still a need for better outcomes because crop production prediction is a difficult task in precision agriculture (Filippi *et al*., 2019).

Predicting agricultural yield has become a highly intricate challenge in the modern era. However, advancements in machine learning have emerged as the most promising method to accurately forecast crop production even before cultivation, without the need for extensive human intervention (Das *et al.*, 2023). By spotting patterns, correlations, and trends in datasets, Machine Learning (ML) makes it possible to extract insightful information from them. These models need to be trained using datasets that depict the results based on prior knowledge. Predictive models are constructed by combining several features; and during training phase their parameters are defined using previous data. A part of data from training phase is saved for model testing and performance evaluation. There are several difficulties in developing a high-performance predictive model in ML research. To successfully address these problems, it is critical to select the appropriate algorithms; and both the platforms and algorithms must be capable of handling huge amounts of data (Zhu *et al.*, 2023). The integration of ML into agriculture marks a significant advancement towards sustainable and data-driven farming practices. It empowers stakeholders with accurate information, reducing uncertainties and maximizing productivity. As technology continues to evolve, the application of ML in agriculture is likely to expand, leading to more efficient and resilient food production systems in future (Alpaydin, 2010).

Various ML methods have extensively been employed for crop yield prediction, leveraging their capability to recognize nonlinear patterns in large datasets. Models such as LSTM, GPR, and Holt-winter time series have been validated for sugarcane yield prediction, with LSTM demonstrating the highest accuracy (Saini *et al.*, 2023). Additionally, RF models calibrated with different predictors exhibited promising results for sugarcane yield forecasting, with an optimal RF model achieving a RMSE of 9.9 t ha^{-1} (Dos Santos Luciano *et al.*, 2021). Integration of ML algorithms, as seen in opti-SAR sugarcane yield prediction model, has shown improved accuracy as compared to the single-base models. Combining ML algorithms with remote sensing has also proved effective, outperforming the individual classification methods and resulting in relatively low mean square error values for various crops (Ilyas *et al.*, 2023). Further, ML techniques such as neural networks (NN), deep learning (DL), random forest (RF), support vector regressions (SVR), and gradient boosting trees (GBT) have widely been used in crop yield modelling, with RF and NN models often demonstrating the highest prediction accuracies (Jagtap *et al.*, 2022). However, challenges arise with small sample sizes in high-dimensional feature spaces, for which Bayesian models and ensemble learning techniques like bagging, boosting, and model stacking provide effective solutions, improving the generalizability of ML models fitted over sparse data (Vabalas *et al.*, 2019).

Everingham *et al.* (2016) have emphasized the urgency of harnessing data mining and ML technologies to enhance industry insights for making critical business decisions that promote sustainable agricultural systems. In their study, the researchers employed RF models on three distinct datasets related to biomass and climate indices. By using these datasets, they successfully forecasted

sugarcane yields, highlighting the potential of ML techniques to provide valuable guidance to the agricultural industry. With the ability to analyse vast amounts of data, data mining and ML can significantly contribute to the informed decision-making, ensuring the long-term sustainability of agricultural practices. With rapid improvements in ML techniques, several researches have studied the utility of these methods in forecasting agricultural outcomes. Several researchers have combined crop model derivatives with ML approaches to create statistical models for yield prediction (Liu *et al.*, 2017).

Machine learning (ML) is not limited to yield prediction alone; but has also been harnessed to forecast variations in rainfall patterns, crucial for agriculture. By assimilating historical meteorological data and employing sophisticated algorithms, ML models can provide valuable insights into the potential changes in rainfall distribution, thereby help the farmers to adapt the practices for optimizing water usage and mitigate the impacts of droughts or excess precipitation (Xing *et al.*, 2018). In India, the use of ML in advanced agricultural yield assessments has recently gained attraction, with research utilizing ML algorithms to forecast sugarcane, paddy, and cotton yields (Guruprasad *et al*., 2019; Prasad *et al*., 2021; Arumugam *et al.*, 2021; Mourão *et al*., 2021). Using moderate-resolution satellite photos that are suitable for analysis, Nihar *et al.* (2022) forecasted the sugarcane crop production for Uttar Pradesh (UP). In UP, district-level sugarcane yield predictions were trained using ML approaches. Further, the accurate crop masks and high-resolution satellite photos are required to obtain satisfactory outcomes. Neethi *et al.* (2023) worked on the efficient estimation of mango canopies yield using ML method while Khan *et al.* (2023) used ML and stochastic pattern analysis for forecasting time-series data. Lovesum *et al.* (2023) analysed best fit ML method for smart agriculture. The current study had two primary objectives *viz*., to analyse historical sugarcane yield data to identify temporal patterns and dependencies using time series modelling, specifically the Autoregressive Integrated Moving Average (ARIMA) method, as well as ML techniques like RF, SVM, and Gradient Boosting Machine (GBM); and to explore the effectiveness of autocorrelation in determining the optimal ARIMA model (Cowpertwait and Metcalfe, 2009; Tunnicliffe, 2016). Additionally, ML algorithms such as RF, SVM, and GBM may be used to estimate sugarcane yield, and compare their predictive capabilities with ARIMA model. The ultimate goal was to identify the most accurate approach for estimating sugarcane yields, thus may provide valuable insights for effective agricultural planning and decision-making concerning India's sugarcane production and trade.

MATERIALS AND METHODS

Data

This study utilized Machine Learning (ML) and time series modelling to forecast sugarcane crop yield in India based on data spanning the past twenty years (2001 to 2020). The year-wise data on sugarcane productivity was extracted from the website National Statistics on Sugarcane, ICAR (*https://sugarcane.icar.gov.in/index.php/sugarcane-statistics/*) and deployed in R for analysis. The data for this study were divided into a training set (80%) and a holdout set (20%) for model assessment.

Time series modelling

Time series analysis plays a pivotal role in forecasting crop yields, enabling experts to analyse historical data over time and identify patterns, trends, and seasonality in production. Among the various methodologies, the seminal work of Box and Jenkins serves as a foundational reference (Box *et al*., 2015). Their approach known as the Box-Jenkins methodology provides a systematic framework for time series modelling, including the development of Autoregressive Moving Average (ARMA) and Auto-regressive Integrated Moving Average (ARIMA) models.

ARMA model: The ARMA model comprises of two components: Auto-regressive (AR) and Moving Average (MA) components. In ARMA (p, q) model, "p" represents the AR component's order, and

"q" signifies the MA component's order. The AR component captures autocorrelation by incorporating lagged values, while MA component relates the present values to past forecast errors. Mathematically:

$$\text{AR component: } y(t) = c + \sum \phi_i y_{t-i} + \epsilon_t$$
$$\text{MA component: } y(t) = c + \sum \theta_i \epsilon_{t-i}$$

Where, *y(t)* is the value of time series at time *t; c* represents the constant or intercept in the model*;* ϕ_i's are the parameters of the autoregressive (AR) part of the model; θ_i's are the parameters of the moving average (MA) part of the model. And ϵ_t represents the error term at time t.

Extending the ARMA framework, ARIMA incorporates an integrated (I) component to address non-stationarity. In ARIMA (p, d, q) model, "p" represents the AR order, "d" signifies the differencing order, and "q" represents the MA order. The integrated component account for the number of differencing steps required to achieve stationarity. Mathematically:

$$\text{Integrated component: } y(t) - y(t-d)$$

Linkage between ARMA and ARIMA models: ARIMA encompasses ARMA, allowing for both stationary and non-stationary data. ARIMA simplifies to ARMA when data is stationary (d = 0), and it employs differencing to handle non-stationarity. Cowpertwait and Metcalfe (2009) further elaborated on these concepts, emphasizing ARIMA's versatility in handling various time series data characteristics.

Machine learning algorithms

Machine learning (ML), a branch of artificial intelligence, utilizes algorithms to learn from data and uncover correlations, providing solutions to various problems. In the realm of crop yield prediction, previous crop year data serves as training data to estimate current crop production. Machine learning encompasses three main types:

a) Supervised ML: This category of algorithms utilizes labelled variables and a mapping function from input to output. If "X" represents the input variable and "Y" represents the output variable, the function can be denoted as Y = f(X). Supervised learning is further divided into regression (for continuous output) and classification (for discrete output) models. The goal is to find a function that accurately maps the inputs to their corresponding targets, minimizing prediction errors.

b) Unsupervised ML: In contrast to supervised learning, unsupervised ML operates solely on input variable "X" without labelled output or target variable "Y." Algorithms in this category aim to discover patterns or structures within the data. For example, K-means clustering is used for grouping data points into clusters based on similarities.

c) Reinforcement ML: Unlike supervised and unsupervised learning, reinforcement learning involves the algorithm interacting with an environment and learning from feedback. The algorithm observes actions and responses to environmental data to maximize prediction accuracy. Deep reinforcement models are applied to optimize agricultural environments for crop improvement.

In summary, ML algorithms, depending on the type, can learn from labelled data (supervised), discover patterns from unlabelled data (unsupervised), or interact with an environment to optimize predictions (reinforcement). These techniques have diverse applications and are essential tools for addressing a wide range of problems in various fields, including agriculture (Pavani and Augusta, 2023).

Random Forest ML algorithm

Random Forests (RF) are robust ensemble learning techniques capable of handling both classification and regression tasks. They construct a group of decision tree models to predict categorical or continuous outcome variables. Each decision tree iteratively divides the data into more homogeneous subsets (nodes) based on predictor variables, enhancing the predictability of response variable. The "RF" tool from the R statistical package was utilized to build random forest classification as well as regression models. To introduce diversity, split points of each tree were randomly selected from a subset of all available predictor variables. For regression models, the default random subset size was one-third of all available predictors, while classification models used the square root of the total

number of predictors. Each node in trees was constrained to a minimum size of one for classification tasks and five for regression tasks to control tree growth and reduce calculation times.

RF technique evaluates the Out-of-Bag (OOB) component of data's regression prediction error to determine the relative importance of predictor variables. This OOB section, representing about 30% of dataset, is not utilized in decision tree construction process. Variable relevance for classification models is determined by mean percent fall in classification rate when a particular variable is eliminated. Conversely, for regression models variable significance is assessed as the average increase in mean squared error caused by the deletion of a specific predictor variable. It is noteworthy that default parameter values of RF package were used in this study for consistency across the models. By leveraging RF and assessing variable importance, valuable insights into the factors affecting the response variable are gained, facilitating better prediction and understanding the sugarcane production dynamics. This method ensures robust and reliable results, enabling data-driven decisions and recommendations for agricultural planning and policy-making.

Support vector machine (SVM): SVM, a powerful ML technique, is used in regression and classification tasks. In this research, SVM was employed for regression to predict sugarcane production, a task involving forecasting a continuous response variable. SVM regression process entails finding the best hyperplane in the feature space that separates data points into different classes, representing various levels of sugarcane production. SVM aims to maximize the margin between data points and hyperplane, enhancing the model's ability to generalize unseen data. The SVM component in R statistical package was used to construct SVM regression models. To achieve optimal performance, the SVM model's hyper-parameters were optimized using the strategies like grid search and cross-validation. In SVM modelling, selecting the appropriate kernel function, such as linear, polynomial, or Radial Basis Function (RBF), is crucial as it affects how the data are transformed into a higher-dimensional space. In SVM regression, the projected value of continuous response variable (sugarcane production) is based on the mean fitted response computed from all individual trees formed from each bootstrapped sample. SVM algorithm aims to generate precise and accurate forecasts by minimizing the error between the anticipated and actual sugarcane production figures.

Gradient boosting machine (GBM): Gradient boosting (GB), an ensemble learning method, is used for regression and classification tasks. In this study, GB was employed to anticipate sugarcane production, focusing on continuous response variable forecasting. GB sequentially assembles a group of decision trees, with each tree built to correct the mistakes of those preceding it, thereby enhancing the overall forecast accuracy. The intricate linkages and patterns found in sugarcane production data are better understood thanks to this sequential learning process. Additionally, the data for this study were divided into a training set (80%) and a holdout set (20%) for model assessment. The GBM package in R was used to create GB models. The model's hyper-parameters, such as maximum tree depth, learning rate, and number of boosting iterations, were optimized using techniques like grid search and cross-validation to obtain the ideal values. GB employs the gradient of loss function to guide the growth of subsequent decision trees, iteratively partitioning the data into more homogeneous subsets. By aggregating predictions from each tree, weighted by learning rate that regulates each tree's contribution, the anticipated value of continuous response variable (sugarcane production) is determined.

Analysis and interpretations

Time series analysis: The Autoregressive Integrated Moving Average (ARIMA) model requires Auto-correlation Function (ACF) and Partial Auto-correlation Function (PACF), two essential analytical approaches used in time series modelling. These functions can be used to assess the statistical significance and linkages between observations in a univariate time series. ACF calculates the relationship between a time series current value and its lag values. In other words, it assesses how the current value Y_t and its previous values $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ relate to one another. One can identify the correlations existing in data by using comprehensive auto-correlation diagram that the ACF gives. On the other hand, PACF delves deeper into the relationships by assessing the correlation of the remaining

effects after removing the variations explained by previous lags. In essence, PACF gives us insights into the correlation of next lag after accounting for the effects of all the earlier lags. By capturing the remaining correlation, the PACF helps to understand the direct relationship between specific lags.

In ARIMA modelling process, the first crucial step is to test the stationarity in time series data. Stationarity is important because ARIMA models work best with stationary data, where the statistical properties like mean and variance remain constant over time. The Augmented Dickey-Fuller (ADF) test is a commonly employed to determine if a time series is stationary or not. If ADF test indicates stationarity, we can proceed with the modelling process. Creating a correlogram, which shows ACF and PACF of stationary time series, comes next after proving stationarity. The PACF assesses the correlation after accounting for the effects of earlier lags, whereas ACF analyses the correlation between the present value and its lagged values. We may choose the correct orders of ARIMA model, such as the autoregressive (AR) and moving average (MA), by analysing the correlogram.

Fitting metrics including the Akaike Information Criterion (AIC), corrected AIC (AICc), and Bayesian information criterion (BIC) were used to find the best-fitted ARIMA model for predicting sugarcane yield. By taking into the account both the model's complexity and goodness of fit, these criteria aid in the selection of models. Better-fitted models with ideal accuracy and simplicity trade-offs have lower values for AIC, AICc, and BIC. With the data transformed into a stationary series, we fitted the ARIMA model so that it captures the relevant temporal patterns and dependencies.

Machine learning methods

Three distinct machine learning algorithms, *viz*., SVM, RF and GBM, each leveraging unique strengths, were employed to predict sugarcane productivity. Data pre-processing preceded model training to optimize the performance. Evaluation metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were utilized to assess the model accuracy so as to guide the selection of the most suitable model for predicting sugarcane productivity from 2021 to 2030. The evaluated models included ARIMA (MAE: 34938.34, MSE: 1346939724, RMSE: 36700.68), RF (MAE: 70299.79, MSE: 6411363140, RMSE: 80070.99), SVM (MAE: 35823.69, MSE: 1997556362, RMSE: 44694.03), and GBM (MAE: 83643.48, MSE: 8548489015, RMSE: 92458.04).

Yield prediction

The yield prediction of sugarcane crop between 2021 and 2030 were computed using diverse modelling techniques like ARIMA, RF, SVM, and GBM. Leveraging data from the website National Statistics on Sugarcane, ICAR (*https://sugarcane.icar.gov.in/index.php/sugarcane-statistics/*), these forecasts provide valuable insights into future productivity trends. By using the methods ranging from time series analysis to advanced machine learning algorithms, analysts sought to offer comprehensive forecasts, aiding the stakeholders in making informal decisions within the agricultural domain.

RESULTS AND DISCUSSION

Time series analysis

The initial ADF test showed that the time series data are not stationary (Table 1). However, further research, including the use of auto.arima function and visual examination of ACF and PACF plots (Fig. 1) revealed that differencing the data causes it to become stationary. This differencing process removes

Table 1: Augmented Dickey-Fuller test

Data: Sugarcane production		
Dickey-Fuller	Lag order	p-value
-2.7262	2	0.2957

Alternative hypothesis: Time series data is stationary

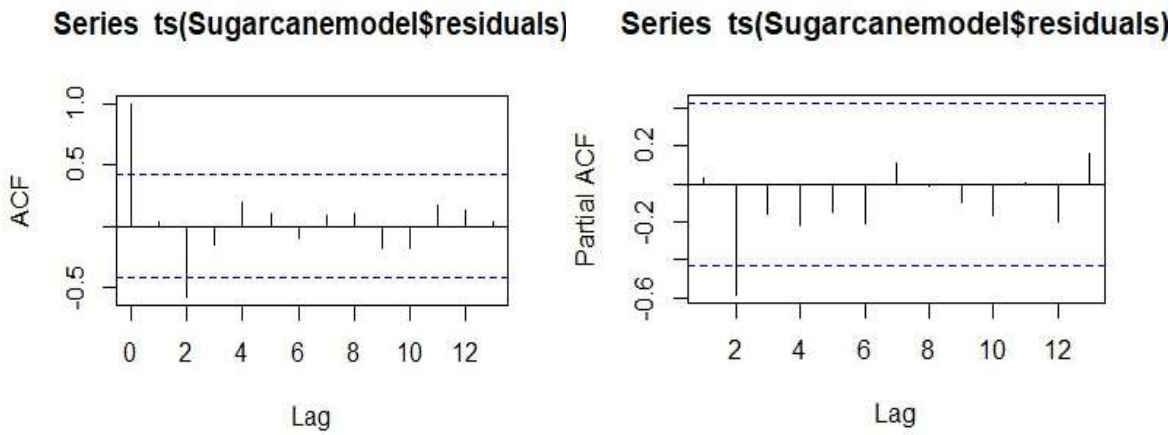

Fig. 1: ACF and PACF plots for sugarcane production time series data

underlying trends or seasonality. With the data transformed into a stationary series, we confidently fitted the ARIMA model, ensuring that it captured the relevant temporal patterns and dependencies. This facilitated the accurate predictions for the time series, such as sugarcane yield in this context. The Table 2 displays the best fitted ARIMA model along with the corresponding fitting measures, such as AIC, AICc, and BIC, which allowed us to compare different ARIMA models so as to select the model that strikes the right balance between accuracy and simplicity.

Table 2: Different ARIMA models with AIC and parameters of best fitted ARIMA model

ARIMA models	Akaike information criteria	For best model ARIMA (0,1,0)	
ARIMA(2,1,2) with drift		sigma^2	1.456e+09
ARIMA(0,1,0) with drift	458.3662	log likelihood	-227.4
ARIMA(1,1,0) with drift	460.3446	AIC	456.79
ARIMA(0,1,1) with drift	459.5727	AICc	457.03
ARIMA (0,1,0)	456.7942	BIC	457.74
ARIMA(1,1,1) with drift	∞	-	-

In time series analysis, particularly for forecasting sugarcane productivity, the ARIMA model stands as a widely used method. This approach incorporates three key components: autoregression (AR), differencing (I), and moving average (MA). To commence ARIMA modelling process, it is imperative to assess the statistical properties and interdependencies within the time series data. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) serve as indispensable tools for this purpose. These functions help in unveiling the degree of correlation between observations at different time lags as well as aid in identifying potential patterns and trends. Further, to ensure the efficacy of ARIMA model, it is essential to address the issue of stationarity within the time series data. Augmented Dickey-Fuller (ADF) test emerges as a prominent method for testing stationarity. If initial assessment reveals non-stationarity, then differencing the data is employed to stabilize the mean and variance, effectively removing any underlying trends or seasonality. Following the stationarity transformation, correlograms are constructed to scrutinize the ACF and PACF. These plots provide insights into the lag structure of time series and assist in determining the appropriate orders of ARIMA model.

To select the most suitable ARIMA model, various fitting metrics such as Akaike Information Criterion (AIC), corrected AIC (AICc), and Bayesian Information Criterion (BIC) are utilized. These metrics offer a quantitative means of evaluating the trade-off between model complexity and goodness of fit. Typically, ARIMA model with lowest AIC, AICc, or BIC values is deemed the best-fitted model. In the context of specific study, after conducting rigorous assessment and model selection procedures, the chosen ARIMA model was identified as ARIMA (0,1,0) based on its superior fitting measures. This model configuration signifies no autoregressive terms, one differencing term, and no moving

average terms. Thus, it captures the essential dynamics of sugarcane productivity time series effectively, paving the way for reliable forecasts and informed decision-making. The findings of this study align with previous research conducted by Mishra *et al.* (2021), who used ARIMA models to explain and forecast the production of sugarcane. This coherence underscores the effectiveness and applicability of ARIMA modelling in the domain of sugarcane production forecasting, consolidating its position as a valuable tool for agricultural research and planning.

Machine learning methods

Three machine learning algorithms *viz*., SVM, RF, and GBM, were applied to predict sugarcane productivity. Data pre-processing preceded model training to optimize the performance. Evaluation metrics for assessing the model accuracy included MAE, MSE, and RMSE. Among the evaluated models, ARIMA demonstrated lowest errors, making it the most promising choice for predicting sugarcane productivity. RF exhibited higher errors, indicating potential inadequacy for this dataset. SVM performed moderately well, closely trailing ARIMA in performance. However, GBM showed highest errors, suggesting limited suitability for this application.

SVM proven most efficient with multi-dimensional data and limited training datasets. In contrast, RF gained favour for its efficient implementation and ability to mitigate overfitting via independent decision rules. Meanwhile, GBM functioning as an ensemble learner progressively refined the models from weaker learners, effectively addressing instances with high mistakes during the learning process. Employing the gradient descent method, GBM minimized errors. Before model training, data pre-processing was conducted to optimize performance. Evaluation metrics such as MAE, MSE and RMSE were utilized to gauge model accuracy, guiding the selection of the most suitable model for predicting sugarcane productivity from 2021 to 2030.

Table 3: Adequacy of different algorithms over holdout data

Models	MAE	MSE	RMSE
ARIMA	34938.34	1346939724	36700.68
Random Forest (RF)	70299.79	6411363140	80070.99
Support Vector Machine (SVM)	35823.69	1997556362	44694.03
Gradient Boosting Machine (GBM)	83643.48	8548489015	92458.04

MEA = Mean absolute error, MSE =Mean square error, RMSE = Root mean square error

Amongst the evaluated models ARIMA (MAE: 34938.34, MSE: 1346939724, RMSE: 36700.68), RF (MAE: 70299.79, MSE: 6411363140, RMSE: 80070.99), SVM (MAE: 35823.69, MSE: 1997556362, RMSE: 44694.03), and GBM (MAE: 83643.48, MSE: 8548489015, RMSE: 92458.04), ARIMA emerged as the most promising choice for the dataset due to its comparatively lower errors. RF exhibited notably higher errors, indicating potential inadequacy for this dataset. SVM performed moderately well, closely trailing ARIMA in performance. However, GBM showed highest errors among the models assessed, suggesting its limited suitability. Therefore, in present study ARIMA appeared as the preferred model for accurate predictions. The findings of this study align with Mourão de Almeida *et al.* (2021) thus focusing on the importance of accurate forecasting in sugarcane production. By employing ARIMA method and evaluating various statistical measures, including AIC, RMSE, MAE, and MAPE, the study identified suitable forecasting models for sugarcane production in India and its major producing states.

Yield prediction

The Table 4 presents detailed comparison of yield predictions (in '000 t) for sugarcane crops in India spanning from the years 2021 to 2030, employing ARIMA, RF, SVM, and GBM models. Notably, ARIMA consistently projected lower yields as compared to the machine learning models. This discrepancy highlights the distinct methodologies and underlying assumptions inherent in each modelling approach.

The visualization of these predictions in Fig. 2 further illustrates the divergence between actual and forecasted sugarcane productivity, providing valuable insights into the performance of each model over the forecasted period. Moreover, these findings resonate with prior research by Sneha and

Bhavana (2023), emphasizing the efficacy of machine learning techniques in agricultural yield prediction. This comparative analysis enhances our understanding of sugarcane productivity dynamics and informs strategic decision-making in the agricultural sector, contributing to improved planning and resource allocation.

Table 4 presents the yield prediction (in '000 t) of sugarcane crop in India for the years 2021 to 2030 using ARIMA, Random Forest, SVM, and GBM. ARIMA consistently predicted lower yields as compared to the other models. These predictions are further visualized in Fig. 2, depicting both the actual and predicted productivity of sugarcane.

Table 4: Yield prediction of sugarcane crop

Year	ARIMA ('000 t)	Random Forests ('000 t)	SVM ('000 t)	GBM ('000 t)
2021	333992.1	391036.2	376383.3	405396.7
2022	333992.1	391036.2	372393.2	405396.7
2023	333992.1	391036.2	367032.3	405396.7
2024	333992.1	391036.2	361027.8	405396.7
2025	333992.1	391036.2	355078.4	405396.7
2026	333992.1	391036.2	349727.0	405396.7
2027	333992.1	391036.2	345297.2	405396.7
2028	333992.1	391036.2	341895.7	405396.7
2029	333992.1	391036.2	339460.4	405396.7
2030	333992.1	391036.2	337829.5	405396.7

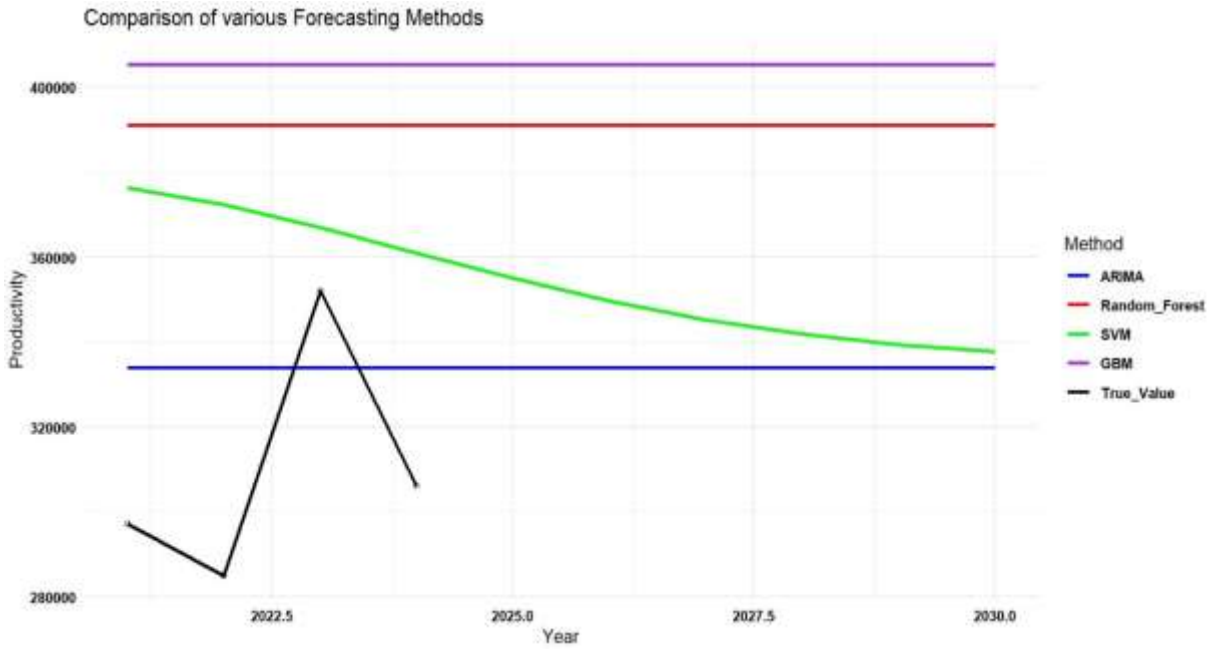

Fig. 2: Plot of comparison of sugarcane prediction (in '000 t) using different forecasting methods

In conclusion, ARIMA emerged as the preferred model for accurate predictions of sugarcane productivity. Its ability to capture temporal patterns and dependencies, along with its comparatively lower errors, makes it well-suited for this application. However, further refinement and evaluation of machine learning models could potentially improve the predictive performance in future studies. The study is subject to the limitations, notably the assumption of stationarity inherent in the ARIMA model, which may not always hold true in real-world agricultural datasets. Additionally, while ARIMA showed promise, its simplicity might overlook nuanced patterns, warranting further exploration of more sophisticated modelling approaches. Furthermore, the influence of external factors like climate

and economic conditions on model performance underscores the importance of thorough validation and consideration of broader contextual factors for reliable predictions.

Conflict of interest: The authors declare no conflict of interest.

Funding information: There is no funding for this work.

REFERENCES

Alpaydin, E. 2010. *Introduction to Machine Learning* (2nd edn.). The MIT Press Cambridge, Massachusetts, London, England.

Arumugam, P., Chemura, A. and Schauberger, B. 2021. Remote sensing based yield estimation of rice (*Oryza sativa* L.) using gradient boosted regression in India. *Remote Sensing*, **13**(12): 1-18.

Bhatt, R. and Singh, P. 2021. Sugarcane response to irrigation and potassium levels in a soil testing low in available potassium. *Agricultural Research Journal*, **58**(4): 709-715.

Bhatt, R., Majumder, D., Tiwari, A.K., Singh, S.R., Prasad, S. and Palanisamy, G. 2023. Climate-smart technologies for improving sugarcane sustainability in India - A review. *Sugar Tech*, **25**(1): 1-14. [*https://doi.org/10.1007/s12355-022-01198-0*].

Box, G.E. P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. 2015. *Time Series Analysis: Forecasting and Control (5th edn.)*. John Wiley & Sons Inc., New Jersey, USA.

Cowpertwait, P.S.P. and Metcalfe, A.V. 2009. *Introductory Time Series with R* (1st edn.). Springer. [*https://doi.org/10.1007/978-0-387-88698-5*].

Das, A., Kumar, M., Kushwaha, A., Dave, R., Dakhore, K.K., Chaudhari, K. and Bhattacharya, B.K. 2023. Machine learning model ensemble for predicting sugarcane yield through synergy of optical and SAR remote sensing. *Remote Sensing Applications: Society and Environment*, **30**: 100962. [*https://doi.org/10.1016/j.rsase.2023.100962*].

Dos Santos Luciano, A.C., Picoli, M.C.A., Duft, D.G., Rocha, J.V., Leal, M.R.L.V. and Le Maire, G. 2021. Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm. *Computers and Electronics in Agriculture*, **184**: 106063. [*DOI: 10.1016/j.compag.2021.106063*].

Everingham, Y., Sexton, J., Skocaj, D. and Inman-Bamber, G. 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for Sustainable Development*, **36**: 27. [*https://doi.org/10.1007/s13593-016-0364-z*].

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D.S.N., Pozza, L.E., Ugbaje, S.U. and Bishop, T.F.A. 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, **20**: 1015-1029.

Guruprasad, R.B., Saurav, K. and Randhawa, S. 2019. Machine learning methodologies for paddy yield estimation in India: A case study. pp. 7254-7257. **In:** *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. Agricultural and Food Sciences, Computer Science Yokohama [*https://doi.org/10.1109/igarss.2019.8900339*].

Ilyas, Q.M., Ahmad, M. and Mehmood, A. 2023. Automated estimation of crop yield using artificial intelligence and remote sensing technologies. *Bioengineering*, **10**: 125. [*https://doi.org/10.3390/bioengineering10020125*].

India. NITI Aayog. 2020. *Sugarcane and Sugar Industry: Final Report of the Task Force. Niti Aayog.* [*https://books.google.co.in/books?id=T3mWzgEACAAJ*].

Jagtap, S.T., Phasinam, K., Kassanuk, T., Jha, S.S., Ghosh, T. and Thakar, C.M. 2022. Towards application of various machine learning techniques in agriculture. *Materials Today Proceedings*, **51**: 793-797.

Khan, A.B.F., Kamalakannan, K. and Ahmed, N.S.S. 2023. Integrating machine learning and stochastic pattern analysis for the forecasting of time-series data. *SN Computer Science,* **4**: 484. [*https://doi.org/10.1007/s42979-023-01981-0*].

Liu, S., Wang, X., Liu, M. and Zhu, J. 2017. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, **1**(1): 48-56.

Lovesum, J. and Prince, B. 2023. A study on experimental analysis of best fit machine learning approach for smart agriculture. *SN Computer Science*, 4: 187. [*https://doi.org/10.1007/s42979-022-01612-0*].

Mishra, P., Al Khatib, A.M.G., Sardar, I., Mohammed, J., Karakaya, K., Dash, A., Ray, M., Narsimhaiah, L. and Dubey, A. 2021. Modeling and forecasting of sugarcane production in India. *Sugar Tech*, **23**(6): 1317-1324. [*https://doi.org/10.1007/s12355-021-01004-3*].

Mourão de Almeida, G., Pereira, G.T., Rabelo de Souza Bahia, A.S., Fernandes, K., Marques Júnior, J. 2021. Machine learning in the prediction of sugarcane production environments. *Computers and Electronics in Agriculture*, **190**: 106452. [*https://doi.org/10.1016/j.compag.2021.106452*].

Neethi, M.V., Kiran, A.G. and Tiwari, H. 2023. Intelligent mango canopies yield estimation using machine vision. *SN Computer Science*, **4**: 171. [*https://doi.org/10.1007/s42979-022-01602-2*].

Nihar, A., Patel, N.R., Pokhariyal, S. and Danodia, A. 2022. Sugarcane crop type discrimination and area mapping at field scale using Sentinel images and machine learning methods. *Journal of the Indian Society of Remote Sensing*, **50**(2): 217-225.

NITI Aayog, India, 2020. Sugarcane and Sugar Industry: Final Report of the Task Force. Niti Aayog. Government of India, New Delhi [*https://books.google.co.in/books?id=T3mWzgEACAAJ*].

Pavani, S. and Augusta, S.B.P. 2023. Improved precision crop yield prediction using weighted-feature hybrid SVM: Analysis of ML algorithms. IETE *Journal of Research.* [*https://doi.org/10.1080/03772063.2023.2192000*].

Prasad, N.R., Patel, N.R. and Danodia, A. 2021. Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research*, **9**(2): 195-206.

Saini, P., Nagpal, B., Garg, P. and Kumar, S. 2023. CNN-BI-LSTM-CYP: A deep learning approach for sugarcane yield prediction. *Sustainable Energy Technology and Assessments*, **57**: 103263. [*https://doi.org/10.1016/j.seta.2023.103263*].

Sneha, V. and Bhavana, V. 2023. Sugarcane yield and price prediction using forecasting models. pp. 1-6. **In:** *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*. [*https://doi.org/10.1109/ICECONF57129.2023.10084094*].

Vabalas, A., Gowen, E., Poliakoff, E. and Casson, A.J. 2019. Machine learning algorithm validation with a limited sample size. *PLoS One,* **14**(11): e0224365. [*https://doi.org/10.1371/journal.pone.0224365*].

Wilson, G.T. 2016. Time series analysis: Forecasting and control. *Journal of Time Series Analysis*, **37**: [*https://doi.org/10.1111/jtsa.12194*].

Xing, L., Li, L., Gong, J., Ren, C., Liu, J. and Chen, H. 2018. Daily soil temperatures predictions for various climates in United States using data-driven model. *Energy,* **160**: 430-440.

Zhu, L., Liu, X., Wang, Z. and Tian, L. 2023. High-precision sugarcane yield prediction by integrating 10-m Sentinel-1 VOD and Sentinel-2 GRVI indexes. *European Journal of Agronomy*, **149**: 126889. [*https://doi.org/10.1016/j.eja.2023.126889*].