Enhanced tobacco yield prediction using spatial information and exogenous variable-driven machine learning models

Abstract

Remote sensing technology has been essential in studying the relationship between tobacco canopy spectral characteristics and biomass yield. This study has been conducted in Garnepudi, Andhra Pradesh, employed satellite imagery obtained between 2015 and 2023 to extract vegetation indices (VI's). Accurately predicting yield is crucial for India's economy. This study investigates the efficacy of various predictive models for tobacco yield forecasting using multiple vegetation indices: NDVI, GNDVI, SAVI, MSAVI, LAI, and LSWI. The models assessed include traditional parametric approaches (ARIMAX, MLR), machine learning techniques (ANN, SVR, RFR), and advanced ensemble methods like XGBoost. The results highlight XGBoost as the most accurate model, consistently delivering the lowest error metrics, including RMSE and MAE, across all vegetation indices. Specifically, XGBoost achieved the best performance with LAI showing RMSE of 86.657, MAE of 58.324, sMAPE of 14.354, MASE of 1.001, and QL of 29.162 respectively. They exhibited lower error metrics, as compare to the statistical and ML models underscoring their effectiveness and potential in tobacco yield prediction. This study highlights the significant role of remote sensing technology in capturing crop development patterns and accurately forecasting tobacco yield, thereby offering valuable insights for agricultural planning and decision-making.

Key words: *Machine learning; vegetation indices; tobacco; yield prediction; XGBoost*

1. Introduction

Tobacco (*Nicotiana tabacum* L.) is one of the most economically significant agricultural crops in the world. India is the 2nd largest producer and exporter after China and Brazil respectively. In India, Tobacco crop is grown in an area of 0.45 M ha (0.27% of the net cultivated area) producing nearly 750 M kg of tobacco leaf. In the global scenario, Indian tobacco accounts for 10% of the area and 9% of the total production (Indian Tobacco Board, Rajahmundry).

Crop yield estimation can be used to help farmers to mitigate production losses during adverse conditions and enhance production under optimal and favourable circumstances (1). Many countries depend on traditional methods conventional techniques of data collection and ground-based field reports for crop yield estimation (2,3). In recent years a variety of mathematical models and machine learning techniques are proposed for estimating yield of various crops (4).

Remote sensing is the acquisition of information about an object or phenomenon without making physical contact with the object. It relies on the use of electromagnetic radiation as an information carrier to collect data about objects or phenomena from a distance (5) Remote sensing techniques has the capability to offer timely and quantitative information about agricultural crops across large regions (6), and various methods have been developed to estimate crop yield (7).

The use of spectral measurements from crops provides valuable information for estimating various crop parameters throughout the growth cycle. Among the parameters that can be estimated are leaf area index (LAI) (8), plant growth, plant density, crop canopy area,

plant population and canopy total nitrogen status (9,10). These measurements are essential for understanding and managing crop health and productivity.

Numerous studies have focused on predicting tobacco yield using various machine learning techniques. Traditional time-series models like ARIMA and its variations, including SARIMA and ARIMAX, have been widely applied to capture the underlying relationships in agricultural data (11). (12) compared ARIMAX and VAR models for rice price prediction in Thailand, finding that ARIMAX offered superior performance. However, the challenges of nonstationary, nonlinear, and noisy data have driven researchers to explore more advanced methods, such as machine learning (ML) and deep learning (DL) models (13). (14) used advanced ML techniques like KNN, DT, SVM, RF and LASSO regression along with VIs for wheat crop yield prediction. (15) highlighted the use of the ARIMAX model to assess and forecast the impact of climatic and hydrological factors on cash crop production in Bangladesh, providing a novel approach to time-series analysis in agriculture. Similarly, Harish applied ML and DL models for price forecasting of essential crops like Tomato, Onion, and Potato (TOP) in major Indian markets. (16) conducted a comparative analysis of time-series models specifically for onion price forecasting, further expanding the application of these models.

This study addresses a crucial gap in current research by focusing on tobacco yield prediction, utilizing both the primary target variable and related VI's as exogenous variables (17). Additionally, our research undertakes an in-depth comparative analysis of various machine learning techniques, including Artificial Neural Networks (ANNs), Support Vector Regression (SVR), Random Forest Regression (RFR) and XGBoost, alongside statistical models such as Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) and Multiple Linear Regression (MLR). The goal of this study is to enhance tobacco yield prediction by integrating machine learning models that incorporate VIs as exogenous factors.

2. Materials and methods

2.1 Study area

This study has been conducted in the Garnepudi village, which is located in the Prakasam district of Andhra Pradesh state, India. The district's geographical coordinates are latitudes of 15° 05′ N and the longitudes of 70° 93′E. figure 1 depicts the selected tobacco growing fields in the Garnepudi village.

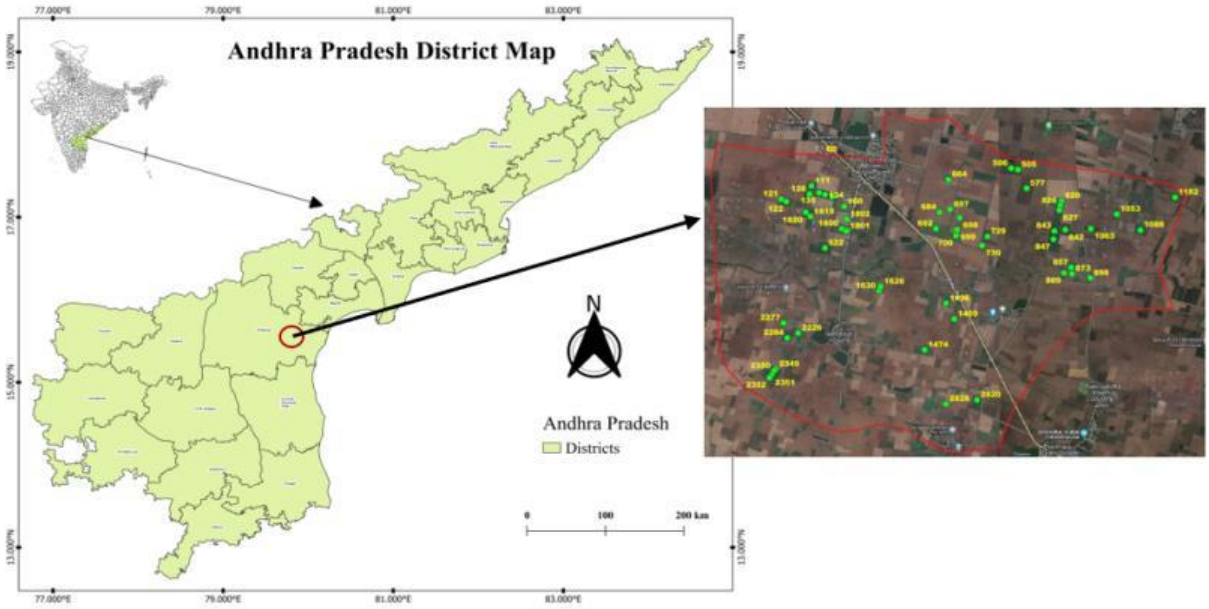


Figure 1: Selected tobacco growing fields in Garnepudi village

2.2 Data description

Experimental data consists of three parts *i.e.* ground truth data, satellite data and ancillary data.

2.2.1 Ground Truth data:

In this study tobacco yield data in time series from 2015-2023 was taken from Garnepudi Rythu Bharosa Kendra (RBK). Data includes yield and latitude and longitude of 55 fields were collected.

2.2.2 Satellite data

Satellite data were derived from Sentinel-2 (S-2) multispectral data, which consist of 13 spectral bands, each with specific wavelength ranges, allowing for detailed analysis and interpretation of the Earth's surface characteristics (https://lta.cr.usgs.gov/sentinel_2). The S-2 data were used for land cover and land use map preparation and for the generation of various vegetation indices. For this study, S-2 images collected from 2015-2023 during crop window period of each year.

2.2.3 Ancillary data:

In addition to the S-2 multispectral data, ancillary data were used, including the district boundary map, village boundary map of Garnepudi which were digitized using ArcGIS software and allowed for precise spatial analysis and integration with other datasets. A village boundary map of Garnepudi in the form of a shapefile was acquired from a survey of India website. (https://onlinemaps.surveyofindia.gov.in/FreeMapSpecification.aspx).

2.3 Methodology for Sentinel-2 Data Extraction

Initially, the ancillary data along with ground truth data were used for the development of region of interest (ROI). Further, the S-2 images were selected from the google earth engine (GEE) catalogue by applying a cloud cover filter, ensuring that only scenes with minimal or no clouds were included. Specifically, a threshold of less than 5% cloud cover was set for the filtering process. The selected S-2 images are from 2015 to 2023, corresponding to the tobacco growing season, which spans from germination to full maturity. After the preprocessing of the S-2 image various vegetation indices were calculated based on the ground truth data.

In this study various vegetation indices are extracted such as normalized difference vegetation index (NDVI), soil-adjusted vegetation index (SAVI), green normalized difference vegetation index (GNDVI), modified soil-adjusted vegetation index (MSAVI), leaf area index (LAI). Figure 2 illustrates the schematic diagram of digital image processing steps for generation of vegetation indices using GEE.

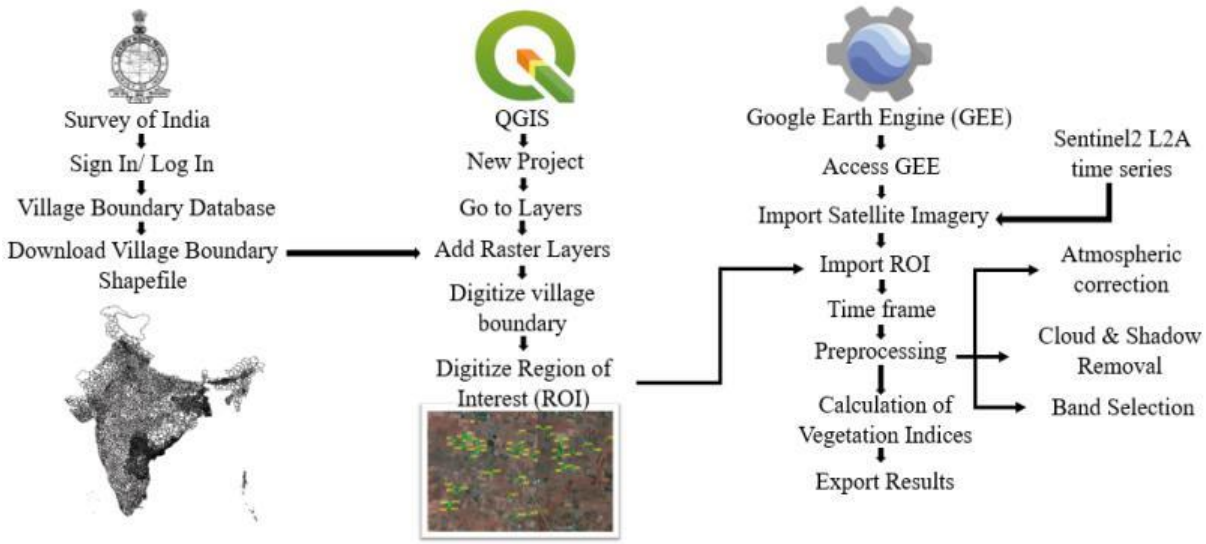


Figure 2: Schematic diagram of digital image processing steps for generation of vegetation

2.3.1 Normalized Difference Vegetation Index (NDVI)

The NDVI is a widely used remote sensing index that assesses vegetation presence and health. First proposed by (18), it is calculated using the reflectance values of the near-infrared (NIR) and red (R) spectral bands. The formula for NDVI is:

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + \mathrm{R}} \quad (1)$$

NDVI values range from -1 to 1, with values close to 1 indicating healthy, dense vegetation, values near 0 indicating bare soil or urban areas, and negative values representing water bodies. It is commonly used to measure crop health and vigour (Figure 3a).

2.3.2 Green Normalized Difference Vegetation Index (GNDVI)

The GNDVI is an enhancement of the traditional NDVI, utilizing the green spectral band along with the NIR band. It is particularly useful for monitoring vegetation health and chlorophyll content of tobacco plants (figure 3b) (19). The formula for GNDVI is:

$$\mathrm{GNDVI} = \frac{(\mathrm{NIR} - \mathrm{Green})}{(\mathrm{NIR} + \mathrm{Green})} \quad (2)$$

where G represents the green band reflectance. GNDVI values range from -1 to 1, with values near 1 indicating robust vegetation health and high chlorophyll content, while values closer to -1 suggest the presence of non-vegetated surfaces or water bodies.

2.3.3 Soil Adjusted Vegetation Index (SAVI)

The SAVI is similar to NDVI but includes an adjustment for the influence of bare soil, making it particularly useful in areas with sparse vegetation. It aims to reduce the impact of soil background on the vegetation signal. The formula for SAVI is:

$$\mathrm{SAVI} = \frac{(\mathrm{NIR} - \mathrm{RED})(1 + \mathrm{L})}{(\mathrm{NIR} + \mathrm{RED} + \mathrm{L})} \quad (3)$$

Here, L is a user-defined parameter, typically set to 0.5, that adjusts for soil background. SAVI values range from -1 to +1, with higher values indicating healthier vegetation. Although SAVI was developed to correct for soil brightness, it can still be sensitive to soil background variations due to the adjustment parameter L (20). It is especially useful in agricultural applications for crop monitoring (figure 3c).

2.3.4 Modified Soil-Adjusted Vegetation Index (MSAVI)

The MSAVI is an improvement of SAVI, designed to further reduce soil background effects and enhance vegetation signal accuracy. It is particularly effective in areas with sparse vegetation (21). The formula for MSAVI is:

$$\mathrm{MSAVI} = \frac{\left(2 \times \mathrm{NIR}+1 - \sqrt{((2 \times \mathrm{NIR} + 1)^2 - 8 \times (\mathrm{NIR} - \mathrm{Red}))}\right)}{2} \quad (4)$$

MSAVI values range from -1 to +1, with higher values indicating healthier vegetation and lower values representing less vegetation or bare soil. MSAVI is advantageous for agricultural applications where soil background can interfere with vegetation indices, providing more accurate estimates of vegetation cover and health (figure 3d).

2.3.5 Leaf Area Index (LAI)

The Leaf Area Index (LAI) measures the total leaf area per unit ground area and is an important parameter for assessing vegetation density and health (22,23). LAI can be estimated from remote sensing data using various models and indices, such as the NDVI or SAVI. The formula for LAI varies depending on the model used, but it generally relates to the amount of vegetation cover:

$$\mathrm{LAI} = \frac{1}{\mathrm{K}} \times \left(\frac{(\mathrm{K} \times (\mathrm{NIR} - \mathrm{Red}))}{(\mathrm{NIR} + \mathrm{Red})} + 1\right) \quad (5)$$

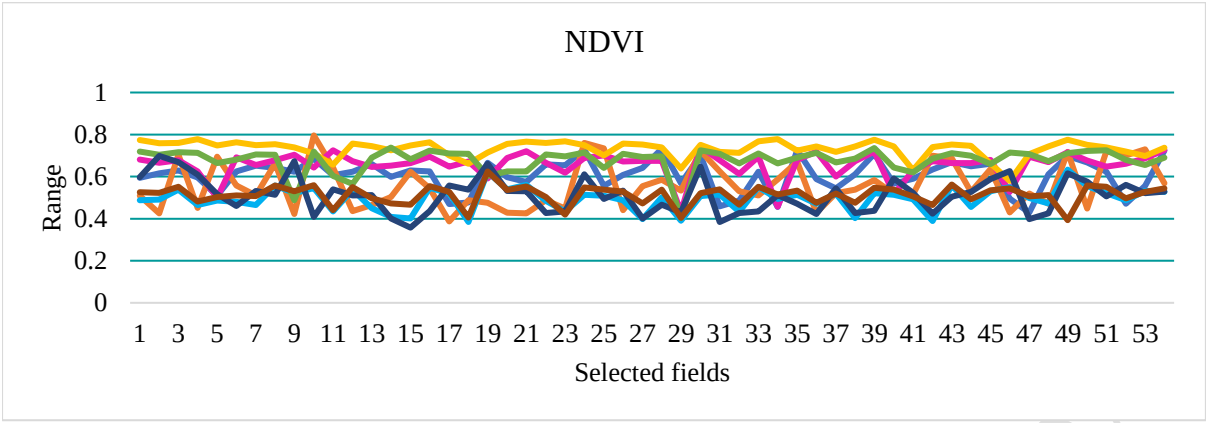
Where K is an empirical constant, usually set to 1.5. Higher LAI values indicate denser vegetation and more leaf coverage, important for predicting crop yields (Figure 3e).

2.3.6 Leaf Surface Water Index (LSWI)

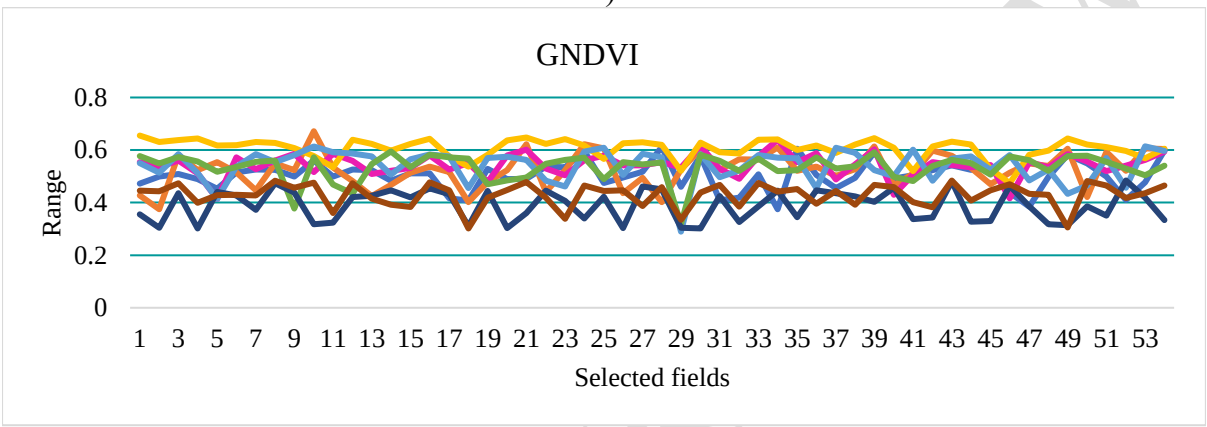
The Leaf Surface Water Index (LSWI) is a vegetation index designed to assess the water content in vegetation, serving as a key indicator of plant water status and overall health. It is particularly sensitive to the presence of water in leaves, making it useful for detecting drought stress or assessing water content during different growth stages (24). LSWI is calculated using the NIR and SWIR bands from remote sensing data:

$$\mathrm{LSWI} = \frac{(\mathrm{NIR} - \mathrm{SWIR})}{(\mathrm{NIR} + \mathrm{SWIR})} \quad (6)$$

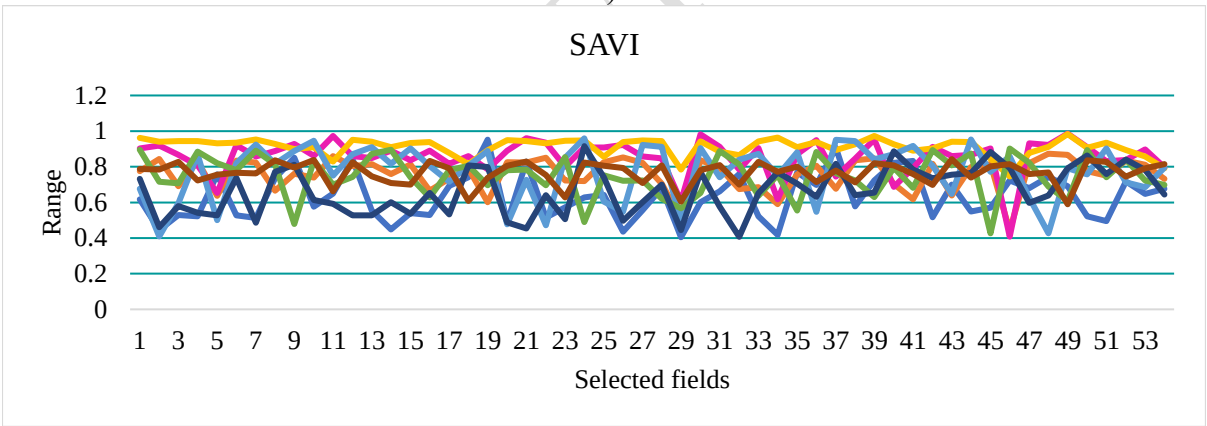
LSWI values typically range from -1 to 1, with positive values indicating higher water content. It is often combined with NDVI for comprehensive crop monitoring, aiding in irrigation decisions and yield predictions. Compared to NDWI, LSWI better captures overall water presence in vegetation, making it particularly useful for hydrological studies and water resource management (Figure 3f).


NDVI
Range
1
0.8
0.6
0.4
0.2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53
Selected fields

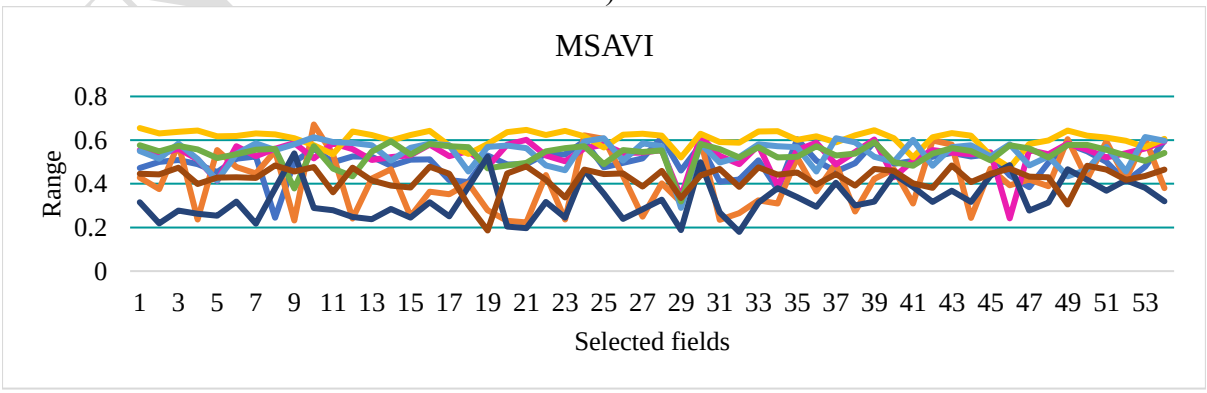
a)


GNDVI
Range
0.8
0.6
0.4
0.2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53
Selected fields

b)


SAVI
Range
1.2
1
0.8
0.6
0.4
0.2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53
Selected fields

c)


MSAVI
Range
0.8
0.6
0.4
0.2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53
Selected fields

d)

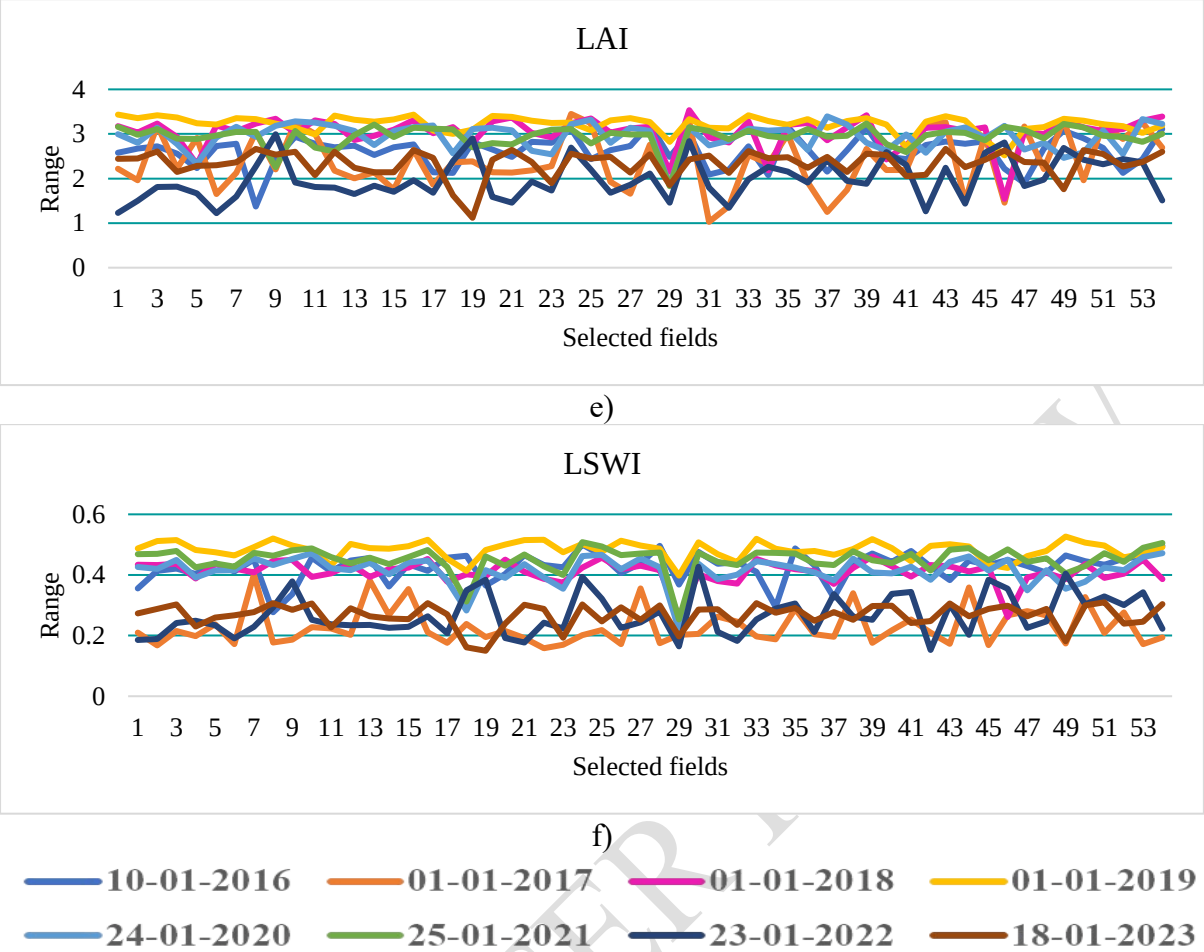


Figure 3: Range of vegetation indices in selected fields in every year

3.4 Methodology for yield prediction

3.4.1 Autoregressive Integrated Moving Average with Exogenous Inputs (ARIMAX)

The ARIMAX model extends the traditional ARIMA model by incorporating external predictors, which improves the accuracy of time series forecasting. The standard ARIMA model, represented as ARIMA $((p, d, q))$, captures temporal dependencies in data, where *'p'* indicates the order of the autoregressive component, *'d'* represents the degree of differencing, and *'q'* signifies the order of the moving average component. ARIMAX enhances this by including exogenous variables, external factors that influence the time series data, making it more versatile for various applications. Mathematically, the ARIMAX $(p, d, q)(P, D, Q)_s$ model can be expressed as:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + X_t \beta + \varepsilon_t \quad (7)$$

Where, Y_t represents the observed value at time t, c is a constant term, $\phi_1, \phi_2, \ldots, \phi_p$ are autoregressive coefficients, $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q}$ are error terms from past time steps, X_t represents the exogenous input variables at time t, β represents the coefficients for the exogenous variables, and ε_t is the error term at time t.

To estimate ARIMAX models, the autoregressive, differencing, and moving average components are fitted to historical time series data, while incorporating exogenous variables. These external inputs enable the model to account for factors outside the immediate time series

that might influence the data, leading to improved forecasting performance (12). ARIMAX models are particularly valuable when there is a discernible temporal pattern in the data, and when additional external variables provide critical information for making more accurate predictions.

3.4.2 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a key statistical technique used to model the relationship between a dependent variable (y) and two or more independent variables $(x_1,\ x_2, \dots, x_n)$. The model is based on the assumption that there is a linear relationship between the dependent variable and the predictors, and it can be represented by the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (8)$$

Where, β_0 represents the intercept, $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the coefficients representing the influence of each independent variable, $x_1, x_2, \dots, x_n$, and ε represents the error term accounting for unexplained variability in the data.

The primary goal of MLR is to estimate the coefficients $(\beta_0, \beta_1, \beta_2, \dots, \beta_n)$ in such a way that the sum of the squared differences between the observed values (y) and the predicted values $\hat{y}$ is minimized.

The coefficients are typically estimated using the method of least squares, which involves minimizing the residual sum of squares (RSS). The RSS is calculated as:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (9)$$

Here, N represents the number of observations. The predicted values $\hat{y}_i$ are determined by multiplying each independent variable $(x_1, x_2, \dots, x_n)$by its corresponding coefficient $(\beta_0, \beta_1, \beta_2, \dots, \beta_n)$, adding the intercept (β_0), and accounting for the error term ε.

MLR is extensively used in various disciplines to explore and quantify the relationships between multiple variables, making it an essential tool for both predictive modeling and data analysis (25).

3.4.3 Artificial Neural Networks (ANN)

ANNs are a type of machine learning model inspired by the structure and function of neurons in the human brain. They are particularly effective for regression tasks, as they can identify and model complex patterns within data. ANNs are composed of layers of interconnected nodes, including an input layer, one or more hidden layers, and an output layer. Each connection between nodes has an associated weight, and each node processes the weighted sum of its inputs using an activation function.

In regression tasks, the output layer usually consists of a single node that represents the predicted continuous value $(\hat{y})$. During the training phase, ANNs adjust their weights through a process called backpropagation. This process involves calculating the error between the predicted output and the actual target values (y) and then updating the weights to reduce this error. The training process typically aims to minimize the Mean Squared Error (MSE), a metric that quantifies the average squared difference between the predicted and actual values:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (10)$$

Where, n represents the number of data points, $\hat{y}_i$ is the predicted value for the i-th instance, and y_i is the actual target value.

A key operation within an ANN involves calculating the weighted sum of inputs (z_i) and applying an activation function (a_i). Common activation functions include the sigmoid function, hyperbolic tangent (*tanh*), and rectified linear unit (ReLU). These functions introduce non-linearities into the model, enabling ANNs to learn and capture complex relationships within the data. The output $(\hat{y}_i)$ of the *i*-th node in the network is determined by:

$$\hat{y}_i = a_i(z_i) \tag{11}$$

ANNs are adept at learning intricate patterns from data, making them well-suited for various regression applications. By adjusting weights and biases during training, ANNs can approximate complex functions, allowing them to accurately model and predict continuous outcomes. Their capacity to capture non-linear relationships makes ANNs a powerful tool for regression analysis across many different fields (17).

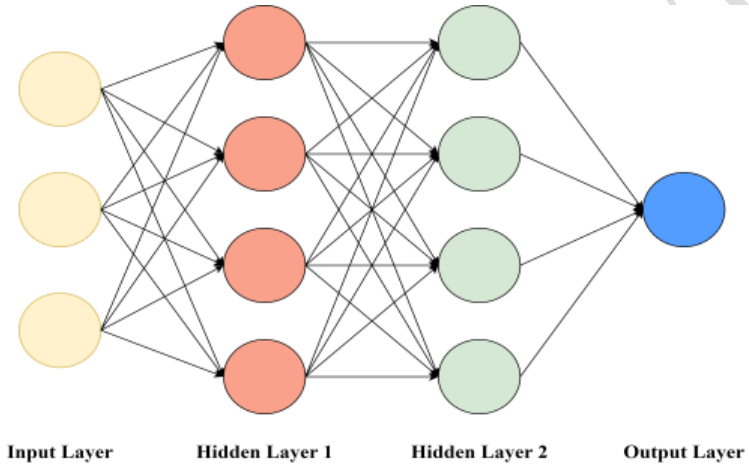


Figure 4: Architecture of ANN

3.4.4 Support Vector Machines (SVM)

SVR is a powerful machine learning algorithm widely used for regression tasks. The core concept of SVR is to find a hyperplane that best fits the data while maximizing the margin, which is the distance between the hyperplane and the nearest data points, known as support vectors. The goal is to minimize prediction errors while allowing for a specified margin of tolerance.

Mathematically, SVR seeks to find a function $f(x)$ that predicts the target values (y) based on input features (y). The objective function for SVR is expressed as:

$$Minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(max(0, |y_i - f(x_i)| - \epsilon)\right)^2 \tag{12}$$

Here, w represents the weights, C is the regularization parameter that controls the trade-off between minimizing the error and maximizing the margin, ϵ is the margin of tolerance, and (x_i, y_i) are the input-output pairs in the training dataset. The function $f(x)$ is determined by the dot product between the input features and the weights, *i.e.*, $f(x) = \langle w, x\rangle + b$, where b is the bias term.

SVR finds the optimal hyperplane by solving a constrained optimization problem, which minimizes errors while striking a balance between fitting the data and maximizing the margin. The final model is influenced by the support vectors—those data points closest to the hyperplane. SVR is particularly effective at capturing non-linear relationships by using kernel functions, which map the input features into a higher-dimensional space where a linear hyperplane can be applied more effectively. This capacity to handle non-linear data patterns makes SVR a versatile tool for regression tasks across a wide range of fields and applications (14,17).

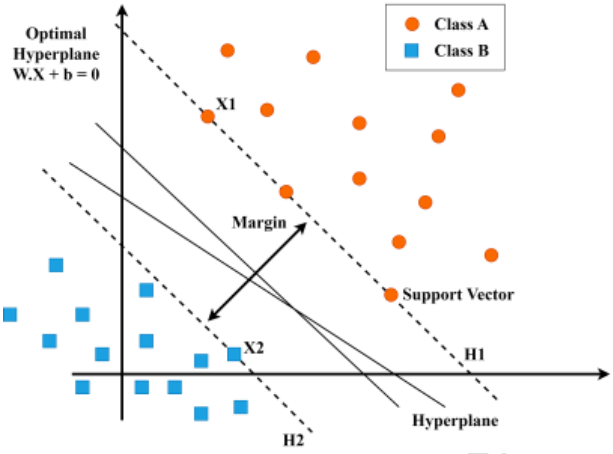


Figure 5: Architecture of SVM

3.4.5 Random Forest (RF)

RF is a regression technique commonly used in data analysis, known for its ability to build an ensemble model by combining the predictive strengths of multiple decision trees. Unlike traditional regression methods, RF constructs multiple decision trees based on an input vector (x) containing various features relevant to the training data. The ensemble model is formed by creating K regression trees and averaging their predictions. The RF regression predictor $\left(\hat{f}_k(x)\right)$ for the input vector x is calculated as follows:

$$\hat{f}_k(x) = \frac{1}{K}\sum_{k=1}^{K} T(x) \tag{13}$$

Here, $T(x)$ represents the individual regression trees grown by RF. To enhance diversity among these trees and prevent correlation, RF employs a technique called bagging. In bagging, training data subsets are created by randomly resampling the original dataset with replacement. This process involves selecting data points from the input sample to generate subsets $\{h(x, \Theta_k), k = 1, \ldots, K\}$, where $\{\Theta_k\}$ are independent random vectors with the same distribution. Some data points may be repeated, while others might not be used, increasing stability and prediction accuracy, especially in the face of slight variations in input data (14,17).

A key advantage of RF is its ability to select the optimal feature or split point from a randomly chosen subset of features for each tree. This approach reduces correlation between trees and helps minimize generalization errors. The RF trees are grown without pruning, which maintains computational efficiency. Additionally, RF uses out-of-bag samples to assess model performance without the need for a separate test dataset. As the number of trees in the forest increases, the generalization error tends to converge, reducing the risk of overfitting.

Furthermore, RF provides valuable insights into the importance of different features, making it a reliable tool for accurate predictions in regression tasks.

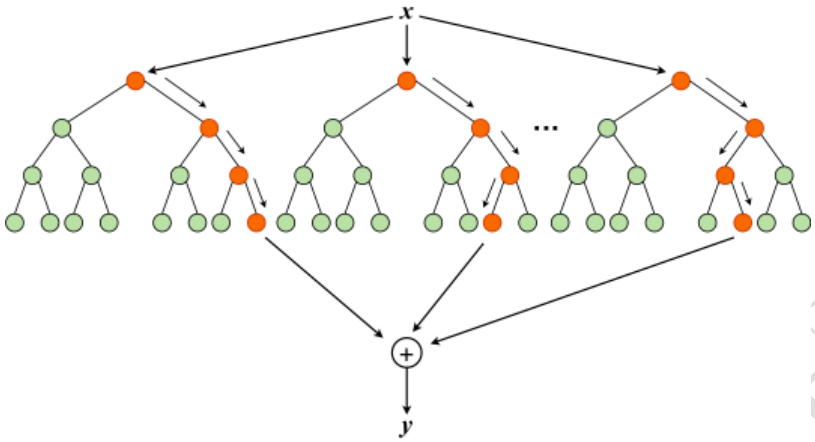


Figure 6: Architecture of Random Forest

3.4.6 Extreme Gradient Boosting (XGBoost)

XGBoost is a sophisticated machine learning algorithm known for its effectiveness in regression tasks. Unlike traditional methods, XGBoost utilizes a gradient boosting framework, which sequentially builds multiple decision trees to improve predictive accuracy. In regression, XGBoost focuses on minimizing an objective function that combines a loss function with a regularization term, ensuring the optimal balance between model fit and complexity. The objective function for XGBoost regression is expressed as:

$$Objective = \sum_{i=1}^{n} \left(\frac{1}{2} . (y_i - \hat{y}_i)^2 + \lambda \, . \Omega(f) \right) \tag{14}$$

Here, y_i represents the actual target value, $\hat{y}_i$ is the predicted value, and n is the number of data points. The term $\Omega(f)$ represents the regularization function, and λ controls the regularization strength.

XGBoost's strength lies in its iterative process. It begins with an initial prediction $\hat{y}_i^{(0)}$ and updates this prediction at each iteration by adding the output from a new decision tree:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{15}$$

Here, t denotes the current iteration, $f_t(x_i)$ is the prediction from the t-th tree for input x_i, and $\hat{y}_i^{(t)}$ is the updated prediction.

To enhance the accuracy of the trees, XGBoost optimizes their structure by selecting the best split points based on the gradient of the loss function. It computes the first-order and second-order gradients for each data point and uses these gradients to determine the optimal splits. Additionally, XGBoost includes a regularization term that manages the complexity of the individual trees, helping to prevent overfitting and improve generalization. By combining the predictions of multiple trees and refining them iteratively, XGBoost produces highly accurate regression models, making it an exceptional tool for a variety of data analysis tasks (26).

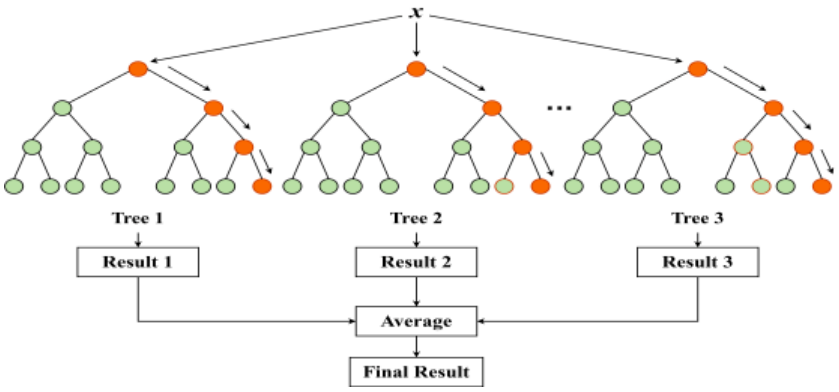


Figure 7: Architecture of XGBoost

3.5 Study Methodology

Creating forecasting models comprises five essential stages: Data Collection, Data Pre-processing, Model Compilation, Model Training, and Model Evaluation. The methodology adopted in this study is depicted through Figure 8. The subsequent section elaborates on these phases in detail.

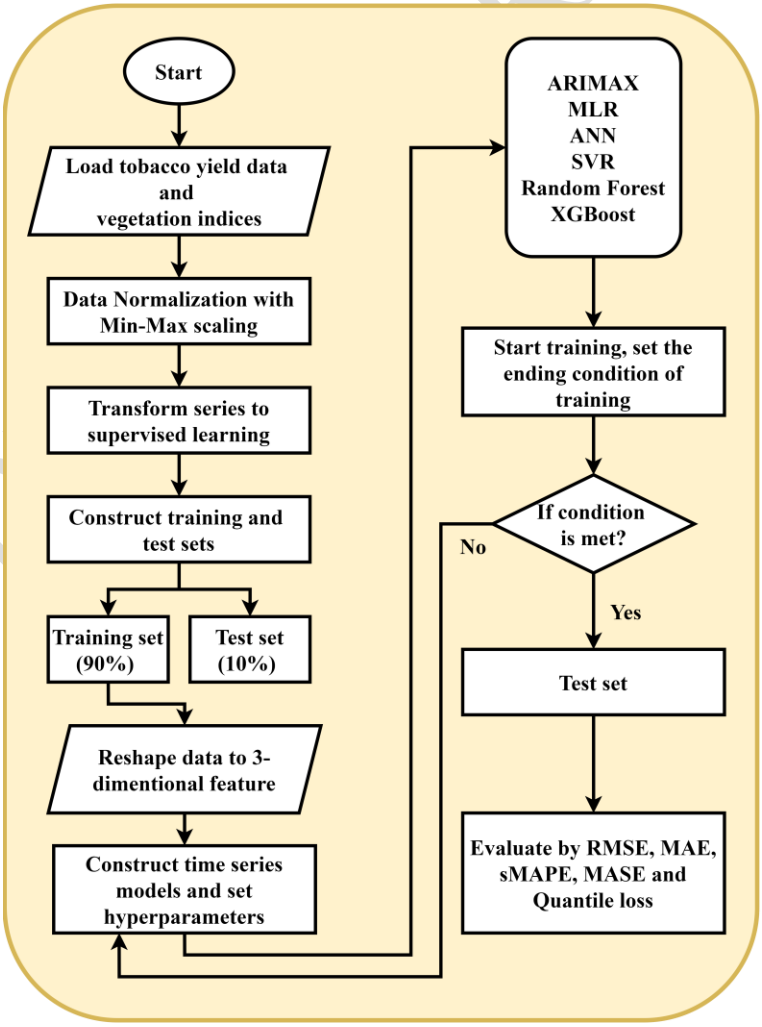


Figure 8: Procedural flowchart of the methodology

3.6 Test for Stationarity

An essential element of time-series analysis is determining whether the data is stationary, meaning that the series has a consistent mean and variance over time. To assess this characteristic, the Augmented Dickey-Fuller (ADF) test was used (27). The findings, as shown in Table 1, offer definitive proof regarding the stationarity of the series.

Table 1: Unit root test results of tobacco

Data	**ADF test**			**Remarks**
	Statistic	p **- value**	**Lags**	
Tobacco	-6.09	0.0007	2	Stationary

3.7 Test for nonlinearity

The Brock-Dechert-Scheinkman (BDS) test, a nonparametric method, was applied to evaluate the presence of nonlinearity in the data series. As shown in Table 2, the probability values calculated within the range of 0.5σ to 2.0σ provide strong evidence supporting nonlinearity in the series, especially for embedding dimensions 2 and 3.

Table 2: BDS test results of tobacco yield

Epsilon	**Embedding dimensions**		p **- value**	**Remarks**
	2	**3**		
0.5σ	255.43	1231.99	< 0.0001	Nonlinear
1.0σ	246.48	770.22	< 0.0001	Nonlinear
1.5σ	201.37	434.72	< 0.0001	Nonlinear
2.0σ	183.85	316.41	< 0.0001	Nonlinear
1.0σ	222.40	646.61	< 0.0001	Nonlinear
1.5σ	219.62	558.85	< 0.0001	Nonlinear
2.0σ	217.94	548.99	< 0.0001	Nonlinear

However, the machine learning models used to analyse agricultural time series data are free from assumptions and excel at efficiently extracting relevant information from time-dependent data.

3.8 Data Pre-processing

Data pre-processing is vital for converting raw data into a format suitable for effective analysis. Through the application of data mining techniques, pre-processing improves the usability and effectiveness of the data, ensuring its quality and reliability for further analysis.

3.8.1 Model building

The development of forecasting models involves two main stages: model training and hyperparameter tuning.

3.8.2 Model Training

The datasets are divided into two segments: a training set and a test set, with a 90:10 split, respectively. After this division, the values within the datasets are normalized to a range of 0 to 1 without altering their distribution, using the following equation:

$$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{16}$$

Here, X_{min}, X_{max} and X_i represent the minimum, maximum, and observed values at time respectively, while X_i' is the rescaled value. The training set, consisting of 90% of the data, captures historical patterns, while the test set uses the remaining 10% for forecasting future points. This structure supports effective model training and evaluation. During training, the target variable (tobacco yield) is modeled alongside exogenous factors like precipitation, temperature, and their interactions to enhance forecasting accuracy. The model is fine-tuned by accounting for these interactions, ensuring comprehensive analysis and robust evaluation.

3.8.3 Model Evaluation

The models were evaluated on the test dataset, comprising the last 10 percent of the complete dataset. Evaluation metrics included Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Symmetrix Mean Absolute Percentage Error (sMAPE), Mean Absolute Scaled Error (MASE) and Quantile Loss (QL).

a) Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2} \tag{17}$$

b) Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}|\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}| \tag{18}$$

c) Symmetric Mean Absolute Percentage Error (sMAPE)

$$\mathrm{sMAPE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{|\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}|}{(|\mathrm{y_i}| + |\hat{\mathrm{y}}_\mathrm{i}|)/2} * 100 \tag{19}$$

d) Mean Absolute Scaled Error (MASE)

$$\mathrm{MASE} = \frac{\mathrm{MAE}}{\mathrm{MAE_{naive}}} \tag{20}$$

e) Quantile Loss (QL)

$$\mathrm{QL_q} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\rho_\mathrm{q}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})\right) \tag{21}$$

where, $\mathrm{q} = 0.5$ and ρ_q is the quantile loss function

$$\rho_\mathrm{q}(\mathrm{e}) = \begin{cases} \mathrm{q.e} & , \text{if } \mathrm{e} > 0 \\ (\mathrm{q}-1).\mathrm{e} & , \text{if } \mathrm{e} \leq 0 \end{cases} \tag{22}$$

where, $\mathrm{y_i}$ is the true values of the variable being predicted, $\hat{\mathrm{y}}$ is the predicted values of the variable and N is the number of observations in the dataset.

4. Results and discussion

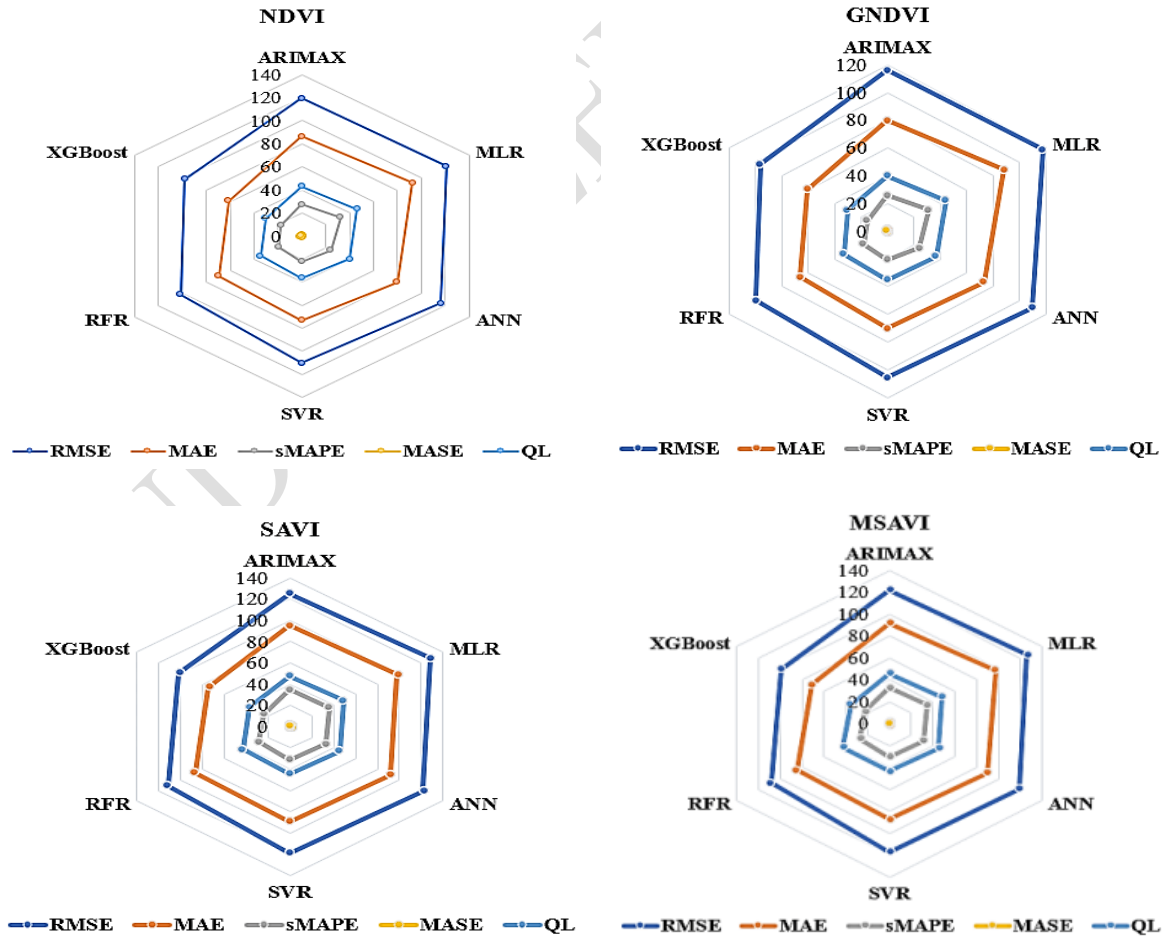
Across various vegetation indices (NDVI, GNDVI, SAVI, MSAVI, LAI, LSWI), XGBoost consistently emerged as the top-performing model. It outperformed other models like ARIMAX, MLR, ANN, SVR, and RFR, achieving the lowest RMSE, MAE, and superior performance in sMAPE, MASE, and QL metrics. For instance, with LAI, XGBoost achieved an RMSE of 86.657, MAE of 58.324, sMAPE of 14.354, MASE of 1.001, and QL of 29.162. This trend was consistent across all indices, making XGBoost the most accurate and reliable model for these analyses as shown in Table 3.

Table 3: Comparative performance metrics of forecasting models for tobacco yield using vegetation indices as exogenous variables

Vegetation indices	Models	RMSE	MAE	sMAPE	MASE	QL
NDVI	ARIMAX	119.568	86.246	27.329	1.576	43.123
	MLR	121.245	92.737	32.446	1.651	46.368
	ANN	116.478	79.909	24.369	1.536	39.954
	SVR	110.244	73.038	22.143	1.519	36.519
	RFR	101.745	69.698	19.089	1.496	34.849
	XGBoost	**98.298**	**61.281**	**17.792**	**1.015**	**30.640**
GNDVI	ARIMAX	115.865	79.784	25.654	1.457	39.892
	MLR	117.542	88.214	30.987	1.587	44.107
	ANN	109.874	73.147	23.941	1.469	36.573
	SVR	105.442	70.126	20.578	1.351	35.063
	RFR	99.547	65.854	18.123	1.119	32.927
	XGBoost	**96.298**	**60.587**	**15.369**	**1.011**	**30.293**
SAVI	ARIMAX	125.357	94.987	34.571	1.479	47.493
	MLR	128.784	98.159	36.147	1.563	49.079
	ANN	121.357	91.357	32.896	1.484	45.678
	SVR	118.871	89.258	31.244	1.472	44.629
	RFR	111.614	86.632	28.687	1.154	43.316
	XGBoost	**101.647**	**74.254**	**23.457**	**1.023**	**37.127**
MSAVI	ARIMAX	122.117	91.842	32.861	1.479	45.921
	MLR	125.971	97.256	34.577	1.563	48.628
	ANN	118.652	90.112	31.344	1.484	45.056
	SVR	117.103	87.141	29.971	1.472	43.570
	RFR	108.492	85.267	27.122	1.354	42.633
	XGBoost	**99.874**	**71.157**	**22.376**	**1.056**	**35.578**
LAI	ARIMAX	111.547	71.457	22.574	1.479	35.728
	MLR	115.159	78.146	24.156	1.563	39.073
	ANN	105.357	70.123	21.456	1.684	35.061
	SVR	101.456	68.143	18.441	1.472	34.071
	RFR	94.789	62.785	16.411	1.154	31.392
	XGBoost	**86.657**	**58.324**	**14.354**	**1.001**	**29.162**
LSWI	ARIMAX	114.547	77.354	24.236	1.479	38.677
	MLR	116.159	82.415	26.232	1.563	41.207
	ANN	109.357	74.896	22.115	1.484	37.448
	SVR	107.456	71.989	20.486	1.472	35.994
	RFR	99.789	69.653	19.987	1.254	34.826

	XGBoost	**91.657**	**61.564**	**16.348**	**1.009**	**30.782**

ARIMAX model can capture temporal patterns, but modelling nonlinear patterns is beyond its capability. Statistical models, in general, are burdened by stringent assumptions that may not always be feasible to satisfy in real-world scenarios. Consequently, ML models such as ANN, SVM, RFR and XGBoost are increasingly favoured due to their data-driven nature and capacity to capture nonlinear patterns. XGBoost offers several advantages that make it a powerful tool in machine learning and data analysis. Its primary strengths lie in its ability to handle large datasets efficiently while offering high accuracy. XGBoost is particularly adept at capturing complex non-linear patterns due to its ensemble learning approach, which combines multiple weak learners to create a strong predictive model. Additionally, it includes built-in mechanisms to prevent overfitting, such as regularization and tree pruning, making it robust in dealing with diverse datasets. The model's flexibility in handling different data types and its scalability to large datasets further enhance its utility across various applications, including time series forecasting, classification, and regression tasks. These characteristics collectively make XGBoost a top choice for many predictive modeling tasks, especially in scenarios where precision and performance are critical. Collectively, the XGBoost strengths of these models significantly boost their performance in prediction of tobacco yield. Additionally, through radar plots (Figure 9) illustrated the performance of the all the ML models for prediction of tobacco yield, where the plots showed a close alignment between actual and predicted values, highlighting the models' robust performance.



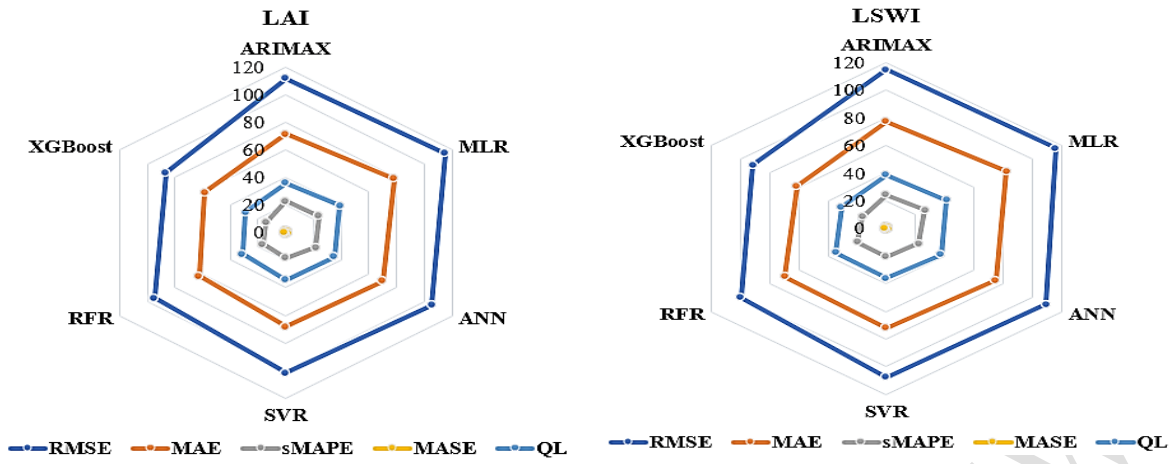


Figure 9: Radar plots for the comparison of performance of different models

Conclusion

In conclusion, the analysis of tobacco yield prediction using various vegetation indices demonstrates that XGBoost consistently outperforms other forecasting models, including ARIMAX, MLR, ANN, SVR, and RFR. XGBoost's superior performance is evident through its lowest RMSE, MAE, and best scores in sMAPE, MASE, and QL metrics across all indices (NDVI, GNDVI, SAVI, MSAVI, LAI, LSWI). This is attributed to XGBoost's ability to handle large datasets, capture complex non-linear patterns, and prevent overfitting through its ensemble learning approach and built-in regularization mechanisms. The model's robustness and flexibility make it a highly effective tool for predictive modeling in tobacco yield forecasting. This research underscores the advantages of advanced machine learning models like XGBoost in agricultural yield prediction, highlighting their capability to manage diverse datasets and deliver precise forecasts. As a result, these findings offer valuable insights for improving predictive accuracy in agricultural economics and provide a solid foundation for future research in this field.

Availability of data and material

In this study tobacco yield data in time series from 2015-2023 was taken from Garnepudi Rythu Bharosa Kendram (RBK). Data will be available based on the request.

Code availability

Code will be available on request to the corresponding author.

Declarations:

Ethics approval, consent to participate, and consent for publication

The manuscript does not report on or involve the use of any animal or human data and "not applicable" in this section.

References

1. Kavita M, Mathur P. Crop yield estimation in India using machine learning. In: 2020 IEEE 5th international conference on computing communication and automation (ICCCA). IEEE; 2020. p. 220–4.

2. Reynolds CA, Yitayew M, Slack DC, Hutchinson CF, Huete A, Petersen MS. Estimating crop yields and production by integrating the FAO Crop Specific Water Balance model with real-time satellite data and ground-based ancillary data. Int J Remote Sens. 2000;21(18):3487–508.

3. Abdul-Jabbar TS, Albayati MMA, Ziboon ART. Environmental factors and wheat crops yield estimation in multiple scales using different remote sensing data/(Al-Zubaydiyah district) as a case study. In: AIP Conference Proceedings. AIP Publishing; 2023.

4. Rani N, Bamel K, Shukla A, Singh N. Analysis of Five Mathematical Models for Crop Yield Prediction. South Asian Journal of Experimental Biology. 2022;12(1).

5. Lillesand T, Kiefer RW, Chipman J. Remote sensing and image interpretation. John Wiley & Sons; 2015.

6. Tao F, Yokozawa M, Zhang Z, Xu Y, Hayashi Y. Remote sensing of crop production in China by production efficiency models: models comparisons, estimates and uncertainties. Ecol Modell. 2005;183(4):385–96.

7. Sishodia RP, Ray RL, Singh SK. Applications of remote sensing in precision agriculture: A review. Remote Sens (Basel). 2020;12(19):3136.

8. Uma Mahesh M, Srinivasan K, Chinnamuthu CR, Shanmugasundaram S, Chandrasekhar CN, Srinivas P. Influence of Different Herbicide Based Weed Management Practices on Growth and Yield of Flue Cured Tobacco in Northern Light Soils of Andhra Pradesh, India. Int J Plant Soil Sci. 2021;33(24):384–90.

9. Jiang Y, Shao X, Li L, Wang T, Zhao H, Hou Q, et al. Remote Sensing Monitoring Model of Tobacco Growth and Yield Based on Ecological Process and Carbon Cycle. J Biobased Mater Bioenergy. 2023;17(2):211–24.

10. Bauer ME. Spectral inputs to crop identification and condition assessment. Proceedings of the IEEE. 1985;73(6):1071–85.

11. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS One. 2018;13(3):e0194889.

12. Anggraeni W, Andri KB, Mahananto F. The performance of ARIMAX model and Vector Autoregressive (VAR) model in forecasting strategic commodity price in Indonesia. Procedia Comput Sci. 2017;124:189–96.

13. Avinash G, Ramasubramanian V, Ray M, Paul RK, Godara S, Nayak GHH, et al. Hidden Markov guided Deep Learning models for forecasting highly volatile agricultural commodity prices. Appl Soft Comput. 2024;158:111557.

14. Naik S, GH HN, Rao SG. Prediction of wheat yield by using UAV RGB drone imagery and advanced machine learning techniques. 2023;

15. Hamjah MA, Chowdhury MAK. Measuring climatic and hydrological effects on cash crop production and production forecasting in Bangladesh using ARIMAX model. Math Theory Model. 2014;4(6):138–52.

16. HT V, MS J, Avinash G, GH HN. A Comparative Analysis of Time Series Models for Onion Price Forecasting: Insights for Agricultural Economics. Journal of Experimental Agriculture International. 2024;46(5):146–54.

17. Nayak GHH, Alam MW, Singh KN, Avinash G, Kumar RR, Ray M, et al. Exogenous variable driven deep learning models for improved price forecasting of TOP crops in India. Sci Rep. 2024;14(1):17203.

18. Rouse JW, Haas RH, Schell JA, Deering DW. Monitoring vegetation systems in the Great Plains with ERTS. NASA Spec Publ. 1974;351(1):309.

19. Shaver T, Khosla R, Westfall D. Utilizing green normalized difference vegetation indices (GNDVI) for production level management zone delineation in irrigated corn. In: The 18th World Congress of Soil Science. 2006.

20. Major DJ, Baret F, Guyot G. A ratio vegetation index adjusted for soil brightness. Int J Remote Sens. 1990;11(5):727–40.

21. Qi J, Chehbouni A, Huete AR, Kerr YH, Sorooshian S. A modified soil adjusted vegetation index. Remote Sens Environ. 1994;48(2):119–26.

22. Yongshi SU, Fuqiang SU, Haibo W, Yanchun C, Jingmin MA, Jun HU, et al. Relationship of spectral characteristics with chlorophyll and leaf area index in flue-cured tobacco leaf. Chin Tob Sci. 2013;34(2):23–7.

23. Zhihong YU, Yong ZOU, Jianjun C, Shiyuan D, Huaiyuan LI, Xiao XU. Spatial-temporal distribution of leaf area index for flue-cured tobacco at mature stage and its correlation with canopy spectral parameters. Tobacco Science & Technology. 2022;55(9).

24. Prabhakar M, Thirupathi M. Remote sensing for pest damage: present status and potential application-case studies from India. 2021;

25. Gopal PSM, Bhargavi R. A novel approach for efficient crop yield prediction. Comput Electron Agric. 2019;165:104968.

26. Ribeiro MHDM, dos Santos Coelho L. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Appl Soft Comput. 2020;86:105837.

27. Kundu MG, Mishra S, Khare D. Specificity and sensitivity of normality tests. In: Proceedings of VI International Symposium on Optimisation and Statistics Anamaya Publisher. 2011.