Original Research Article

Forecasting Green Chilli Prices: Using Data Analytics to Gain Market Understanding

ABSTRACT

Green chilli is a commercially significant vegetable crop grown year-round due to its high demand for both nutritional and health benefits. India stands as the largest producer and consumer of chilli globally, with West Bengal leading in area under cultivation and ranking sixth in production. This study aims to compare and identify the most accurate model for forecasting green chilli prices in the Haldibari market of Cooch Behar district, West Bengal. Price data from January 2015 to May 2024, sourced from AGMARKNET, was used for model development, with 85% of the data allocated for training and 15% for validation. The models tested include the traditional Seasonal Auto-Regressive Integrated Moving Average (SARIMA) and machine learning models like Artificial Neural Networks (ANN) and Support Vector Regression (SVR). Among these, the ANN model is found to be the most accurate, with a low Mean Absolute Percentage Error (MAPE) of 0.14. The model is used to forecast prices for the next 12 months, up to May 2025. The study's findings can aid farmers and policymakers in creating effective crop planning strategies, helping to boost local farm income.

Keywords: Price, Forecasting, SARIMA, ANN, SVR

1. INTRODUCTION

Green chilli is the key ingredient in various cuisines and a rich part of our global agricultural history (Smith, B. D., 1995) [1].The scientific name of green chilli is *Capsicumannum*, which belongs to the Solanaceae, the nightshade family. India is famous for its variety of chilli production.There are four hundred (400) different types of chilli cultivated in India. India plays a major role in chilli consumption. Two million metric tons of chilli has been produced by India in the year 2023 [2]. Andhra Pradesh occupied the first position in chilli production in fiscal year 2023 with 627 thousand metric tonsfollowed by Telangana and Madhya Pradesh with 550 and 320 thousand metric tons respectively[3]. In West Bengal, chilli is cultivated in an area of 65,930 hectares, approximately 27,650 metric tons of production is recorded in West Bengal contributing 6% to total production [4]. Even though it is relatively small compared to major chilli-producing states, it is still meeting its local demands. Coastal areas of West Bengal like South 24-Parganas and Purba Medinipur are the best places for chilli cultivation. Chilli farming has great potential in this state with good agricultural practices.

Price forecasting is a critical component in the economic management of a nation, particularly in the agricultural sector where prices can fluctuate across various markets within a country [5]. This study aims to predict the future prices in selected market. Chilli can be grown year-round due to its greater demand for food value and health benefits. We can suggest that farmers cultivate the crop at a specific time by analysing the best price-meeting months. It is helpful for farmers to produce good output and to increase their economic conditions. This prediction and model accuracy is useful for the stakeholders from farmers to consumers helping them plan and strategize in an unpredictable market. This forecasting provides valuable insight for Agricultural planning, Risk management, Market analysis, Investment decisions, Economic stability, Export and Import decisions, Policy formulations etc.,

In this paragraph, statistical models and techniques used for the study are explained. For forecasting the price, the SARIMA model in traditional models, ANN(Artificial Neural Network) and the Support Vector Regression (SVR) model in machine learning are employed for model fitting and analysis by using several packages in the software R. Different researchers have worked on forecasting commodity prices.Priya *et al*. [6]predicted gold price in India using the SARIMA model.After analysing the data SARIMA (0,1,0) (1,0,1) model was found to be the best fit for forecasting. Vinay *et al.* [7] evaluated ARIMA, SARIMA, BATS, and TBATS models to predict onion prices in Karnataka. Their analysis revealed that the TBATS model outperformed the others, and it was subsequently employed for forecasting future onion prices.Ajith *et al.* [8] studied application of statistical and machine learning models in combination with stepwise regression for predicting rapeseed and mustard yield in Northern districts of West Bengal using a basic linear regression model, machine learning models like Support Vector Regression (SVR), and Artificial Neural Network (ANN).Rathod *et al.* [9] employed artificial intelligence techniques to model and forecast oilseed production in India. Their study compared the performance of ARIMA model, non-linear models like Time Delay Neural Network (TDNN) and Support Vector Regression (SVR). Results from their study revealed that machine learning models dominated the ARIMA model with low RMSE, MAE, and MAPE values.Pavithra et al. [10] applied Exponential Smoothing models and the ARIMA model to analyse potato prices in the Kolar market of Karnataka, they specifically used the Holt-Winter's Exponential Smoothing (H-WES) model for price forecasting.

2. METHODOLOGY

2.1 Data Description

This study is focused on the chilli prices ofHaldibari market in the Northern part of West Bengal state. Secondary data on the monthly prices of green chilli has been collected for the above-mentioned market from the AGMARKNET officialwebsite(https://agmarknet.gov.in/). A total of 9 years and 5 months data is collected from January 2015 to May 2024 with 113 data points for conducting the analysis.

2.2 Methods

2.2.1 Seasonal Autoregressive Integrated Moving Average (SARIMA)

If time series data have seasonal components, then the SARIMAmodel is more useful.The seasonal period is mentioned using an S term at the end of the Seasonal ARIMA model. ARIMA (p, d, q) $(P, D, Q)_{[S]}$.

SARIMA model is represented by the equation:

$(1-\varphi_p B)(1-\Phi_P B^s)(1-B)(1-B^s)y_t=(1-\theta_q B)(1-\Theta_Q B^s)\varepsilon_t$

Where,

B is the backshift operator,

s is the seasonal lag,

ε_t is the sequence of error ~ N (0, σ^2),

Φ and φ represent the seasonal and non-seasonal autoregressive parameters respectively,

Θ and θ represent the seasonal and non-seasonal moving average parameters.

2.2.2 Artificial Neural Network

It contains the input layer which is the first layer, that receives the raw input data, hidden layers are intermediate layers between input and output layers where computations are performed. These layers can be multiple and are not directly observed and output layer is the final layer that produces the result of the network.It also contains **Weights**that act as parameters that connect neurons between layers. Each connection has an associated weight that is adjusted during training to optimize the network's performance. **Biases** are additional parameters in each neuron that allow the activation function to shift to the left or right, aiding in the learning process.It also has an **Activation Function** which is a mathematical function applied to the neuron's output to introduce non-linearity.

The equation for a single hidden layer of neural network is given as:

$$y_{t+1} = g\left(\sum_{j=0}^{q} \alpha_j f\left(\sum_{i=0}^{p} \beta_{ij} y_{t-i}\right)\right)$$

Where,

y_{t+1} = ln (y_{t+1}/ y_t) is the predicted value for y_t at timet,

α_j (j=0, 1, 2, . . ., q) and β_{ij} (i=0, 1, 2,. . ., p; j=1, 2,. . ., q) are the model parameters, also called as the connection weights,

p is number of input nodes, q is the number of hidden nodes,

f and $g$indicates the activation function at hidden and output layer respectively and

$y_{t-i}$is the i[th] input lag of the model.

It includes **Forward Propagation** where input data is passed through the network layers, with each neuron computing a weighted sum of inputs plus a bias term, followed by the activation function. Mean Squared Error (MSE) was used as the loss function for regression tasks, while Cross-Entropy Loss was used for classification tasks.

2.2.3 Support Vector Regression

In this study, three kernels namelyLinear, Radial Basis Function (RBF), and polynomial, were evaluated. The Linear kernel function is suitable for linearlyseparable data and hyperparameters to be optimized while training the SVR model with linear kernel are Cost (c) and Epsilon (ε) while nonlinear kernels like Radial Basis Function (RBF) and polynomial are appropriate for data that is not linearly separated. In the case of the nonlinear RBF kernel function, an additional parameter, Gamma (γ), must also be optimized. For the polynomial kernel function, another parameter to be optimized is the degree of the polynomial (d).

The goal is to find a function f(x) that approximates the target valuesy within a certain margin of tolerance (epsilon). The functionf(x) in SVR is typically represented in the **Table 1**as:

f(x) = W * ϕ(x) + b, where w is the weight vector, ϕ(x) is the mapping function that transforms the input x into a higher-dimensional space, and b is the bias term.

Table 1.Functions of the SVR kernels

Sl. No.	Kernel	Function
1	Linear	$K(x_i, x_j) = x_i x_j$
2	Radial	$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$

3	Polynomial	$K(x_i, x_j) = (X_i X_j + r) d$

Here, RBF kernel (γ) determines the influence of a single training example. Degree (d) is a polynomial Kernel that represents the degree of the polynomial. Coefficient (r) is a polynomial kernel that denotes a constant term that helps in controlling the complexity of the model.

2.3 Accuracy Measures

2.3.1 Akaike Information Criterion (AIC)

The formula for AIC is: AIC = 2k- 2(ln) L

Where,

k is the number of parameters in the model,

L is the maximum value of the likelihood function for the model.

A lower AIC Value indicates a better model, balancing goodness of fit and model complexity.

2.3.2 Root Mean Squared Error (RMSE)

The formula for RMSE is: $\sqrt{\frac{\sum(Y-\hat{Y})^2}{N}}$

2.3.3 Mean Absolute Percentage Error (MAPE)

The formula for MAPE is: $\frac{\sum\frac{|Y-\hat{Y}|}{Y}}{N}$ x 100

Where,

Y are the actual values,

$\hat{Y}$ are the predicted values,

N is the number of observations.

Lower RMSE and MAPE values indicate better model performance.

3. RESULTS AND DISCUSSION

A comprehensive dataset of monthly green chilli prices from the Haldibari market in Cooch Behar district, West Bengal, spanning from January 2015 to May 2024 (113 observations), is divided into training and testing sets. The training set, comprising 85% of the data (96 observations), is used for model development, while the remaining 15% (17 observations) was reserved for testing and evaluating model performance using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) as key metrics.

3.1 Descriptive Statistics

The descriptive statistics for green chilli prices in the Haldibari market are presented in Table 2, based on 113 observations. The average price of green chilli during this period was ₹3,973 per quintal, with prices ranging from a low of ₹857 to a high of ₹9,416. Notably, the data exhibited significant variability, with a Coefficient of Variation (CV) of 46%. The distribution of prices is also characterized by moderate positive skewness and platykurtic tendencies, indicating asymmetry in the data.

Table 2. Descriptive Statistics of green chilli prices in the Haldibari market

Statistics	Price (Rs/Qtl)
Observations	113
Mean	3973.61
Median	3866.37
Minimum	857.60
Maximum	9416.66

Standard Deviation	1852.28
Coefficient of Variation	46.61
Skewness	0.55
Kurtosis	-0.12

The average monthly price of green chilli in Haldibari market, depicted in Figure 1, exhibits seasonality. A clear trend emerges, with prices peaking between July and October, indicating a high-demand period. Conversely, the months of May and June exhibit significantly lower prices, suggesting a period of relative surplus or reduced demand.

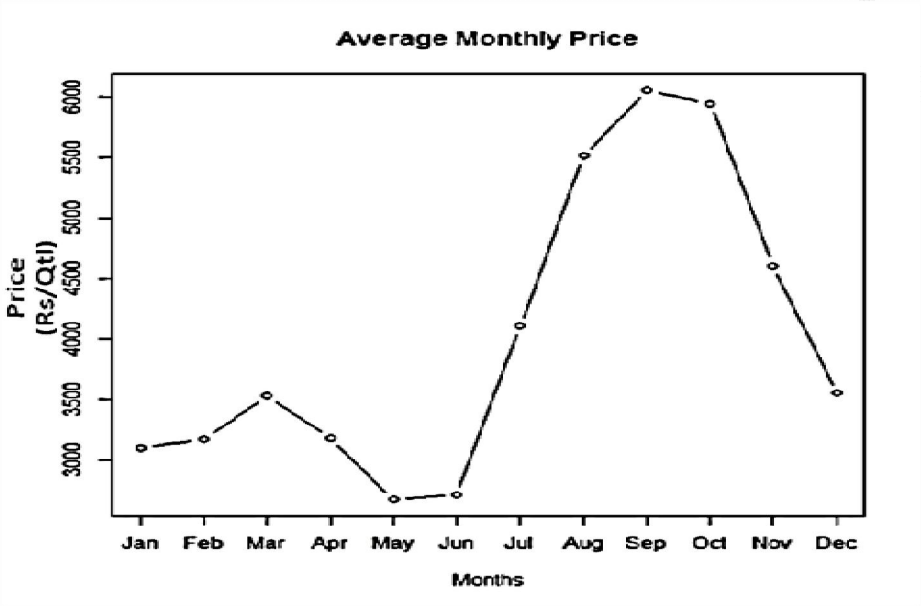

Fig. 1. Monthly average prices of green chilli in the Haldibari market

3.1.1 Seasonal Indices

Seasonal indices are obtained using the simple average method. From the given Table 3, it is observed that the months from July to November have seasonal index values greater than 1, indicating that these months experience seasonality within the year.

Table 3. Seasonal Indices of green chilli prices in the Haldibari market

Months	Seasonal Index
January	0.77
February	0.79
March	0.88
April	0.79
May	0.67
June	0.67
July	1.02
August	1.37
September	1.51
October	1.48
November	1.15

December	0.89

3.2 Results of SARIMA

This seasonality is checked by Kruskal Wallis test as shown in **Table 4**. As the null hypothesis is rejected with a p-value <0.05, the data shows seasonality. Stationarity is checked using the Augmented Dickey-Fuller (ADF) test that results in data found stationarity with the null hypothesis is rejected as the p-value is 0.01(<0.05)as indicated in **Table 5**. The best model ARIMA(2,0,0)(2,0,0)$_{[12]}$ is identified using a low AIC value of 1587.47 and the parameter estimates of the model are presented in **Table6.**

The Ljung-box test is employed to check the residuals of the fitted model for independence. The results found that the p-value is 0.62>(0.05) indicatesindependence of residuals. Accuracy measures like AIC, RMSE, and MAPE for training and testing datasets are taken for the best model which has the lowest values and are presented in **Table 7.**Similar analysis was conducted by Vinay *et al.*[11].

Table 4. Results of the seasonality test for prices in the Haldibari market

Test Used	Test Statistic	p-value
Kruskal-Wallis	42.63	1.25 e^{-05}

Table 5. Results of the stationarity test for prices in the Haldibari market

Test Used	Test Statistic	Lag order	p-value
ADF test	-4.75	4	0.01

Table 6. Parameter estimates of the ARIMA (2,0,0) (2,0,0) $_{[12]}$ model

Model	Coefficient	Estimate	SE	p-value
ARIMA (2,0,0) (2,0,0) $_{[12]}$	AR1	1.0452	0.1108	$2.2e^{-16}$
	AR2	-0.2606	0.1096	0.01
	SAR1	0.1674	0.1199	0.04
	SAR2	0.4433	0.1199	0.01

Table 7. Accuracy measures of the best SARIMA model

Model	AIC	BIC	RMSE		MAPE	
			Training	Testing	Training	Testing
ARIMA(2,0,0) (2,0,0) $_{[12]}$	1587.47	1601.92	846.99	886.81	0.19	0.18

3.3 Results of ANN

To apply non-linear machine learning models, non-linearity is checked using the Brock-Dechert-Scheinkman testas mentioned in the **Table 8.** which results in p-value $2.2e^{-16}$<(0.05)indicates the presence of non-linearity.

Different combination of input, hidden and output nodes are tried in the analysis. Input nodes are varied from 1 to 20, hidden nodes are varied from 1 to 10 using the sigmoid activation function and 1 output node is applied to the data. Out of these combinations, a model with 3 input nodes, 9 hidden nodes and 1 output node is found to be better with low RMSE and MAPEas presented in **Table 9.**Devra *et al.*[12] had also varied different number of input and hidden nodes in their analysis.

Residual analysis is the most important diagnostic measure of the model and itis carried out using the Ljung-Box test that results in a p-value of 0.68 >(0.05) as indicated in **Table 10**. It indicates null hypothesis is accepted, so residuals are linear and independent.

Table 8. Results of the non-linearity test

Test Used	Test Statistic	p-value
BDS test	38.61	$2.2\ e^{-16}$

Table 9. ANN models for green chilli prices in the Haldibari market

Model	RMSE		MAPE	
	Training	Testing	Training	Testing
3-3-1	781	1046	0.16	0.18
3-4-1	718	1291	0.12	0.17
3-5-1	629	1193	0.14	0.19
3-6-1	594	979	0.13	0.15
3-7-1	579	1524	0.12	0.19
3-9-1	**508**	**822**	**0.11**	**0.14**
3-10-1	466	2357	0.12	0.39

Table 10. Results of the residual test of the ANN model

Test Used	Q Statistic	Lag order	p-value
Ljung-Box test	7.39	10	0.68

3.4 Results of SVR

SVR model is fitted to the green chilli prices of Haldibari market using different kernels like Radial basis function, linear and polynomial kernels. The hyperparameters are tuned with the help of caret package by taking values from 0.25 to 1 for cost function, 0.001 to 0.1 for gamma function and 1 to 3 for degree function according to the requirement of the kerneland best combination of the parameters for each kernel is shown in **Table 11.**Here, all kernels are compared with RMSE of training and testing datasets to avoid overfitting of the model and also with MAPE values for finding the best-fitted model with high accuracy. SvmLinear model is found to be best with close RMSE values and less MAPE value of 0.16as shown in **Table 12.**Padipati *et al.*[13] had also conducted analysis using different kernels.

Table 11. Description of SVR model for green chilli prices in the Haldibari market

Sl. No.	Kernel	Cost	Gamma	Degree
1	**Linear**	**1**	**-**	**-**
2	Radial	1	0.03	-
3	Polynomial	1	0.1	1

Table 12. SVR model accuracy measures for green chilli prices in the Haldibari market

Kernel	RMSE		MAPE	
	Training	Test	Training	Test
Linear	**662**	**831**	**0.12**	**0.16**

Radial	706	890	0.13	0.18
Polynomial	680	858	0.15	0.17

3.5Comparison of selected models to identify the best fit

To identify the best-fitting model for the Haldibari market, SARIMA, ANN, and SVR models are compared based on their performance on the testing dataset using RMSE and MAPE metrics. The ANN model emerged as the best model, achieving the lowest RMSE (822) and MAPE (0.14), as shown in **Table 13**. Consequently, the ANN model is selected for forecasting future green chilli prices.

Table 13. Comparison of RMSE and MAPE values for selected models

Market	Variable	Model	RMSE values	MAPE values
Haldibari	Green chilli Price	ARIMA (2,0,0) $(2,0,0)_{[12]}$	886	0.18
		NNAR (3-9-1)	**822**	**0.14**
		Svm Linear	831	0.16

3.6 Forecasting of Monthly Green Chilli Prices

NNAR(3-9-1) model of the Artificial Neural Network(ANN) is used to forecast the prices for 12 months from June 2024 to May 2025 as shown in **Table14** and **Fig2.** The results indicate that July, August, and September show an increase in prices of green chilli when compared to other months.

Table14.Forecasted values of chilli based on the best model

Month	Forecasted price (Rs/Qtl) of chilli using NNAR(3-9-1) model	Month	Forecasted price (Rs/Qtl) of chilli using NNAR(3-9-1) model
June-2024	7739.42	December-2024	5707.08
July-2024	8117.86	January-2025	6227.38
August-2024	7580.73	February-2025	6056.62
September-2024	6323.42	March-2025	5192.34
October-2024	4879.71	April-2025	4343.65
November-2024	5108.50	May-2025	4629.82

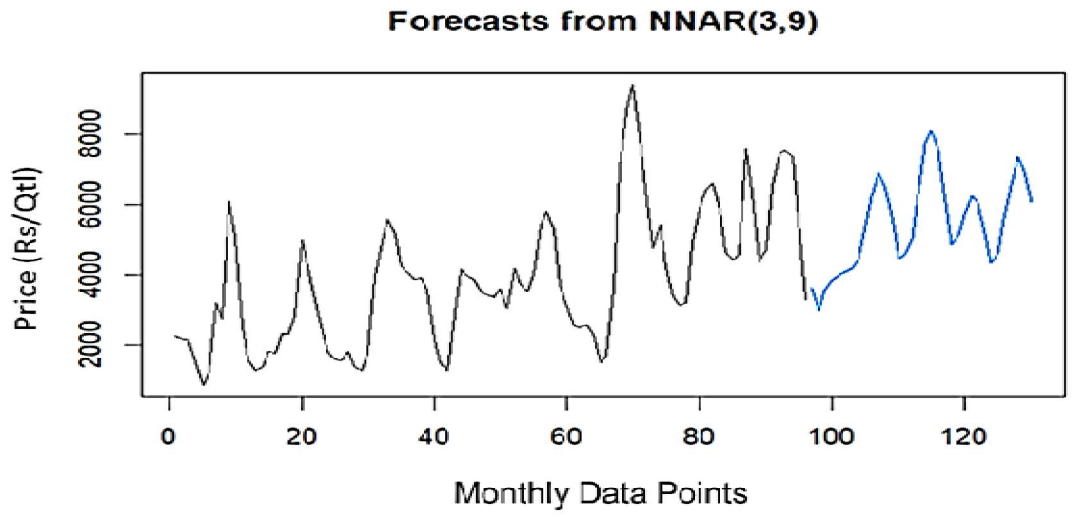

Fig. 2.Graph showing a forecast of green chilli price in the Haldibari market

4.CONCLUSION

Forecasting chilli prices is essential due to the imbalance between production and demand, which significantly impacts market prices. To predict future prices in the Haldibari market, Cooch Behar district, West Bengal, models like SARIMA, ANN, and SVR were evaluated. The ANN model outperformed the others, achieving the lowest RMSE (822) and MAPE (0.14), and was subsequently used for price forecasting in the selected market. The study's findings are valuable for both farmers and policymakers, aiding in effective crop planning. Farmers can focus on increasing chilli production during June, July, August, and September, as these months are forecasted to offer the best prices. Policymakers can also leverage these predictions to better manage the profits and losses of both producers and consumers.

4.1 Limitations and Future scope

In this study, Heteroscedastic time series models are not used as the data taken for the analysis had no volatility in prices. So, the models like Auto Regressive ConditionalHeteroscedasticity (ARCH) and Generalized Auto Regressive Conditional Heteroscedasticity (GARCH) can be used for the data with volatility as observed in the analysis conducted by Sahu *et al.*[14]. Deep learning models like Recurrent Neural Networks (RNN) are also not used in this study but there is a scope to employ those models for the large data sets with more data points. Kumari *et al.* [15] employed the models used in the present study and also the models mentioned above in their research that contains more data.

Disclaimer (Artificial intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

REFERENCES

1. Smith B D. The Emergence of AgricultureScientificAmerican Library. 1995.
2. National Horticulture Board (2023). "Indian Horticulture Database." NHB.
3. APEDA (2023). "Chilli Production in India: 2023 Overview." Agricultural & Processed Food Products Export Development Authority, Ministry of Commerce and Industry, Government of India.
4. Government of West Bengal, Department of Agriculture (2023). "Agriculture Statistics 2023." WB Agriculture Department.
5. Vinay HT, Pavithra V, Jagadeesh MS, Avinash G,and Harish Nayak GH. A Comparative Analysis of Time Series Models for Onion Price Forecasting: Insights for Agricultural Economics. Journal of Experimental Agriculture International. 2024; 46(5): 146-154.
6. Priya S K and Kausalya N. Predicting gold price in India using SARIMA model. International Research Journal of Modernization in Engineering Technology and Sciences. 2024; 6(2): 2587-2594.
7. Vinay HT, Pavithra V, NishthaPradarshika Rai, and Jagadeesh MS. Modelling and forecasting of onion prices in Belgaum market of Karnataka. Int J Agric Extension Social Dev. 2024;7(3S):140-145.
8. Ajith S, Debnath M K, Gupta D S, and Basak P. Application of statistical and machine learning models in combination with stepwise regression for predicting rapeseed-mustard yield in Northern districts of West Bengal. International Journal of Statistics and Applied Mathematics. 2023; 8(3): 141-149.

9. Rathod Santhosha, Singh K N, Patil S G, Naik, R H, Ray Mrinmoy,and Meena V S. Modeling and forecasting of oilseed production of India through artificial intelligence techniques. Indian J. Agric. Sci. 2018; 88(1): 22-27.
10. Pavithra V, Vinay H T, and Soumya Ch. A statistical approach to forecasting potato price in Kolar market of Karnataka.International Journal of Statistics and Applied Mathematics. 2024; 9(2): 32-36.
11. Vinay H T, Basak P, Ghosh A, Ojha S, and Sahu C R. Forecasting of Tomato Price in Karnataka using BATS Model. Journal of the Indian Society of Agricultural Statistics. 2024; 78(2):107-113.
12. Devra S J, Patel D V, Shitap M S and Raj S R. Time series forecasting of price for oilseed crops by combining ARIMA and ANN.International Journal of Statistics and Applied Mathematics. 2023; 8(4): 40-54.
13. Paidipati K K, Chesneau C, Nayana B M, Kumar K R, Polisetty KK, and Kurangi C. Prediction of rice cultivation in India - Support vector regression approach with various kernels for non-linear patterns. AgriEngineering. 2021: 3(2): 182-198.
14. Sahu C R, Basak S, and Gupta D S. Modelling long memory in volatility for weekly jute prices in the Malda district, West Bengal. International Journal of Statistics and Applied Mathematics. 2023; 8(3): 118-124.
15. Kumari P, Goswami V V, and Pundir R S. Recurrent neural network architecture for forecasting banana prices in Gujarat, India. Plos one. 2023; 18(6): e0275702.