RESEARCH ARTICLE

A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh

Mst Noorunnahar[1☯], **Arman Hossain Chowdhury**[2☯], **Farhana Arefeen Mila**[3*]

1 Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh, **2** Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh, **3** Department of Agribusiness, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh

☯ These authors contributed equally to this work.
* famila@bsmrau.edu.bd

Citation: Noorunnahar M, Chowdhury AH, Mila FA (2023) A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. PLoS ONE 18(3): e0283452. https://doi.org/10.1371/journal.pone.0283452

Editor: Sathishkumar V. E., Jeonbuk National University, REPUBLIC OF KOREA

Received: July 26, 2022

Accepted: March 1, 2023

Published: March 27, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0283452

Data Availability Statement: All necessary codes and data are available on GitHub (https://github.com/Arman-Hossain-Chowdhury/Rice-production).

Abstract

In this study, we attempt to anticipate annual rice production in Bangladesh (1961–2020) using both the Autoregressive Integrated Moving Average (ARIMA) and the eXtreme Gradient Boosting (XGBoost) methods and compare their respective performances. On the basis of the lowest Corrected Akaike Information Criteria (AICc) values, a significant ARIMA (0, 1, 1) model with drift was chosen based on the findings. The drift parameter value shows that the production of rice positively trends upward. Thus, the ARIMA (0, 1, 1) model with drift was found to be significant. On the other hand, the XGBoost model for time series data was developed by changing the tunning parameters frequently with the greatest result. The four prominent error measures, such as mean absolute error (MAE), mean percentage error (MPE), root mean square error (RMSE), and mean absolute percentage error (MAPE), were used to assess the predictive performance of each model. We found that the error measures of the XGBoost model in the test set were comparatively lower than those of the ARIMA model. Comparatively, the MAPE value of the test set of the XGBoost model (5.38%) was lower than that of the ARIMA model (7.23%), indicating that XGBoost performs better than ARIMA at predicting the annual rice production in Bangladesh. Hence, the XGBoost model performs better than the ARIMA model in predicting the annual rice production in Bangladesh. Therefore, based on the better performance, the study forecasted the annual rice production for the next 10 years using the XGBoost model. According to our predictions, the annual rice production in Bangladesh will vary from 57,850,318 tons in 2021 to 82,256,944 tons in 2030. The forecast indicated that the amount of rice produced annually in Bangladesh will increase in the years to come.

Introduction

There has been a fast expansion in the world population, which has put a strain on the agricultural sector [1]. Rice is considered the world's third most common major crop, with more than

Funding: The author(s) received no specific funding for this study.

Competing interests: The authors have declared that no competing interests exist.

50% of the world's population eating it as a staple diet [2, 3]. As one of the most nutrient-dense grains, rice is an excellent source of carbohydrate as well as vitamins (B, E, thiamine) and minerals (Ca, Mg, Fe) [4]. About 160 million Bangladeshis rely on rice as a basic meal for their daily diets and survival [5]. Bangladesh's economy is heavily dependent on rice production, which means that the price of rice has a considerable impact on GDP growth, inflation, wages, employment, food security, and poverty [6]. The rice industry employs over 48% of the rural population, provides two-thirds of all caloric intake, and accounts for half of the average person's protein intake [7]. For agricultural GDP and national income, the rice subsector alone contributes about 4.5% to the GDP [8]. Nearly all farming households in Bangladesh cultivate rice. It is produced on about 10.5 million hectares of land, which occupies about 75 and 80% of the total cropped and irrigated areas, respectively [9].

Accurate and timely estimates of crop production before harvest are essential for food security and administrative planning, especially in the current, ever-changing global environment and international scenario [10, 11]. Rice yield forecasting has been extensively examined using various methods all around the world. In order to forecast rice yield, Kumar and Kumar (2012) added fuzzy values to the time series [12]. Alam et al. (2018) applied two hybrid approaches including ARIMAX-ANN and ARIMAX-SVM for estimating rice yield in India [13]. Jing-feng (2011) used NOAA/AVHRR data to predict rice production in Zhejiang Province through ratio models and regression models [14]. Using a crop growth model, Yun (2003) forecasted regional rice production in South Korea [15]. Koide et al. (2013) employed precipitation hindcasts from one uncoupled general circulation model (GCM) and two coupled GCMs to examine the predictive abilities of retrospective seasonal climate forecasts (hindcasts) customized to Philippine rice production data [16]. A satellite remote sensing technique was used by Noureldin et al. (2013) to forecast the production of rice in Egypt [17]. However, to reveal the growth pattern and make the most accurate prediction of rice production in Bangladesh, it is necessary to use a suitable approach that can successfully describe the observed data. Different techniques have been taken to accurately estimate yield, and each method has its own strengths and limitations [18]. For example, Rahman (2010), Mahmud (2018), Rahman et al. (2016), and Sulatana and Khanam 2020 applied the autoregressive integrated moving average (ARIMA) and artificial neural network (ANN) for predicting rice production in Bangladesh [19–22].

Sensor technologies, big data, the Internet of Things, artificial intelligence (AI), and machine learning approaches have recently shown great potential to advance precision agriculture and obtain accurate predictions [23]. According to the aforementioned literature and to the best of the author's knowledge, XGBoost is a machine learning algorithm that has not been widely deployed. The eXtreme Gradient Boosting (XGBoost) model is a supervised machine learning technique and an emerging machine learning method for time series forecasting in recent years [24, 25]. It is a novel gradient tree-boosting algorithm that offers efficient out-of-core learning and sparsity awareness. XGBoost is a supervised learning technique that ought to be particularly good for the problem of claim prediction with both big training data and missing values, even if the commonly used methods such as random forest and neural networks can handle missing values [26, 27]. The robustness of XGBoost results in increased usage of the method in many other applications. As an example, Aler et al. utilize XGBoost in the field of direct-diffuse solar radiation separation by creating two models [28]. Moreover, in infectious disease prediction such as COVID-19, the XGBoost achieved greater prediction accuracy [29, 30].

In contrast, the Autoregressive Integrated Moving Average (ARIMA) model developed by Box and Jenkins (1990) is most widely used for forecasting time series data because of its capacity to handle non-stationary data [31]. The ARIMA model is a suitable forecasting

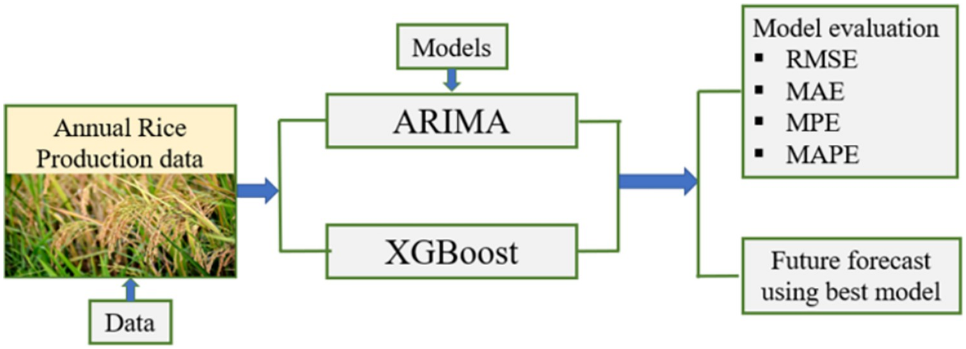

Fig 1. Theoretical framework for the study.

https://doi.org/10.1371/journal.pone.0283452.g001

method in agriculture for different crops and has been extensively used in the fields of economics and finance [31–33]. Therefore, this study aimed to (a) compare the predictive accuracy of the autoregressive integrated moving average (ARIMA) and eXtreme gradient boosting (XGBoost) for accurate modeling the annual rice production data in Bangladesh; and (b) carry out the best model to forecast rice production for the next 10 years (Fig 1). Finally, the findings of this study will help government officials and development practitioners make more accurate short-term predictions of future rice production to boost administrative planning and ensure food security.

Materials and methods

Data source

The annual rice production data from 1961 to 2020 (60 years) used in this study were collected from the website of FAOSTAT [34]. The data were divided into training and test sets. The proportion of training and testing data was 90% and 10%, respectively. The ARIMA and XGBoost models were built using the training data sets. The test data were used to evaluate the predictive ability of the developed models. The data set does not contain any missing values.

ARIMA model

The autoregressive integrated moving average (ARIMA) is a technique for analyzing and predicting time series data that was initially introduced by Box and Jenkins in 1976 [35]. An ARIMA (p, d, q) time series model consists of its three components. The letters p of the ARIMA model denote the autoregressive (AR) order, d denotes the differencing order, and q denotes the moving average order (MA) [36, 37]. The autoregressive order AR(p) describes the linear combination of the observations that are p times earlier with the random shock term, which can be mathematically defined as

$$Y_t = C + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \emptyset_3 Y_{t-3} + \emptyset_4 Y_{t-4} \ldots\ldots \emptyset_p Y_{t-p} + \varepsilon_t \qquad (1)$$

Where, Y_t and ε_t represent the observed value and the random shock terms at time t, $\emptyset_i$ (i = 1,2,3,4. . ..) indicates the model parameters, and c is the constant term. On the other hand, the moving average order MA(q) explains the dependent variable for previous random shock terms, which can be defined as

$$Y_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (2)$$

where, μ represents the mean of the series, θ_j (j = 1, 2, 3. . . q) denotes the model parameters, and q indicates the model's order [38]. According to the above explanation, the ARMA (p, q) model can be defined mathematically as follows:

$$Y_t = C + \mu + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \emptyset_3 Y_{t-3} + \emptyset_4 Y_{t-4} \ldots\ldots + \emptyset_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \theta_4 \varepsilon_{t-4} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3)$$

The general form of the ARIMA (p, d, q) model with the differenced series may be defined mathematically as follows:

$$y'_t = c + \emptyset_1 y'_{t-1} + \emptyset_2 y'_{t-2} + \ldots + \emptyset_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4)$$

Where y'_t explains the difference between the series (the number of differences can be greater than 1);; $\emptyset_1$, $\emptyset_2 \ldots \emptyset_p$ indicate the coefficients of AR(p) terms and θ_1, $\theta_2 \ldots \theta_q$ show the coefficients of the moving average, MA(q) term. More information regarding ARIMA model can be found in the literature [30, 39].

XGBoost model

The eXtreme Gradient Boosting (XGBoost) is a type of boosting application that combines several learning applications to produce higher prediction accuracy than any of the individual learning applications used in several fields [24]. It is a decision tree-based ensemble machine learning approach that is frequently employed in data science. After utilizing an internal approach that aggregates the outcomes from several individual trees, precise forecasts can be obtained [29]. XGBoost was first introduced by Chen Tianqi and Carlos in 2011, and since then several researchers have refined and enhanced it for the follow-up study [40]. The XGBoost model aims to execute a gradient descent optimization approach so that the loss function can be reduced [41]. Boosting is an ensemble technique that can assemble thousands of forecasting models with lower performance into a strong, high-performance model by repeatedly merging the models within permissible parameter values [40, 42]. The objective function can be written as follows:

$$obj(\theta) = \sum_i L(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (5)$$

As mentioned above, the objective function (5) consists of a loss function denoted by L and a regularization term $\Omega(f_k)$, that reduces the new tree's output variation. $\hat{y}_i$ denotes the predicted value and y_i represents the observed value. A detailed information regarding the XGBoost model can be found in the literature [24, 39].

Evaluation parameter of models

One of the major criteria of model evaluation is the calculation of model accuracy. The accuracy of a model describes how the actual and predicted values are close to each other. Model accuracy can be calculated by using several measures [43]. This study used the four widely used model accuracy measures, such as mean absolute percentage error (MAPE), mean percentage error (MPE), mean absolute error (MAE), and root mean square error (RMSE). These measures can be defined mathematically as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \quad (6)$$

$$MPE = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{y}_i - y_i}{y_i} \right) \times 100\% \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{8}$$

$$MAPE = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{9}$$

Where n indicates the number of samples, $\hat{y}_i$ denotes the predicted value and y_i represents the observed value, and $\hat{y}_i - y_i$ indicates the error value. The MAPE measurement provides the percentage result of the errors. Better fitting results are achieved with less errors [41].

Statistical analyses

ARIMA and XGBoost predictive models and several statistical analyses were carried out using RStudio (Version 4.2.1) [44]. The ARIMA model was fitted using the "forecast" package [45]. The XGBoost model was constructed with the "forecastxgb" package. The "ggplot2" package was used for graphical visualization. All necessary codes and data are available at https://github.com/Arman-Hossain-Chowdhury/Rice-production.

Results

The highest amount of rice produced in Bangladesh was 54,905,891 tons in 2020, and the lowest was 13,304,520 tons in 1962. The average amount of rice produced annually in Bangladesh is 29,960,847.08 tons. And the boxplot indicates that the data have no outliers (Fig 2).

We plotted the time series of the annual rice production data from 1961 to 2020 in Bangladesh. The data vary considerably and show a linear pattern. The Augmented Dickey Fuller (ADF) test confirmed that the data are not smooth (Fig 3).

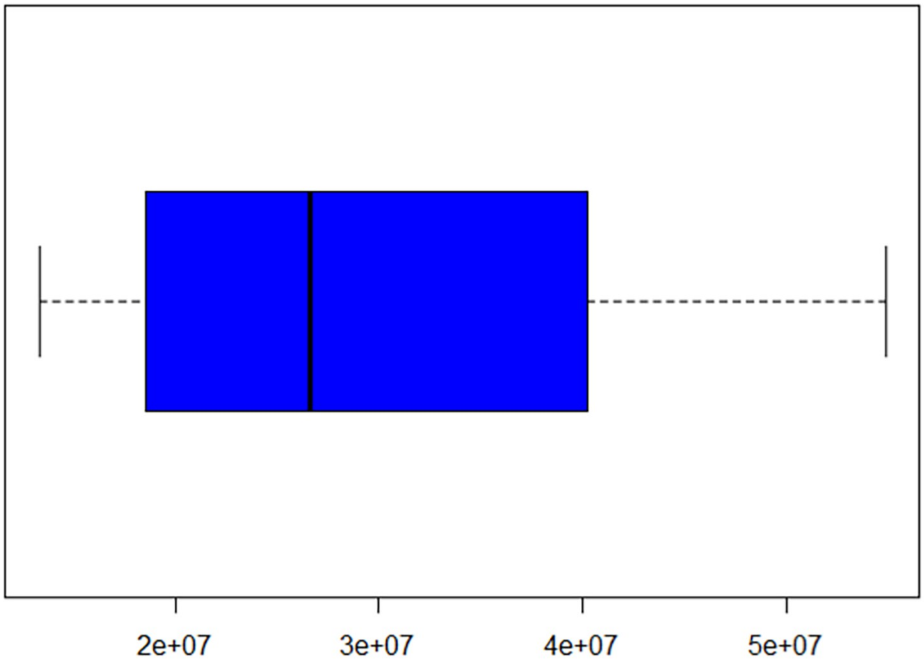

Fig 2. Boxplot of the annual rice production data in Bangladesh from 1961 to 2020.

https://doi.org/10.1371/journal.pone.0283452.g002

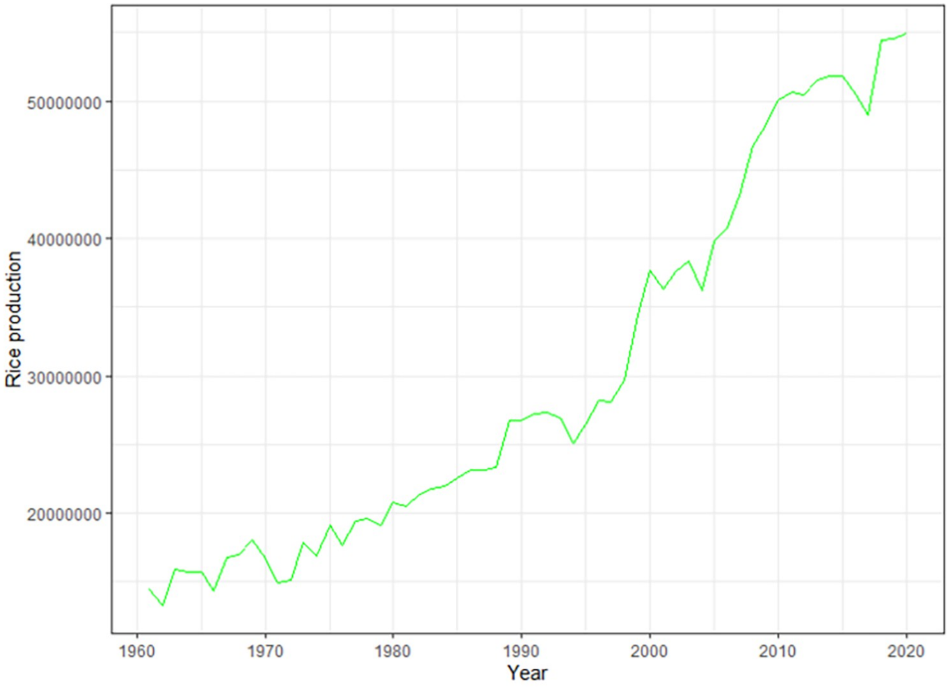

Fig 3. A time series plot for rice production in Bangladesh from 1961 to 2020.

https://doi.org/10.1371/journal.pone.0283452.g003

To reduce variation and stabilize the actual data, Box & Cox (1964) presented a parametric power transformation technique [46]. We applied this technique to make the data stable and exhibit less variation (Fig 4) [47].

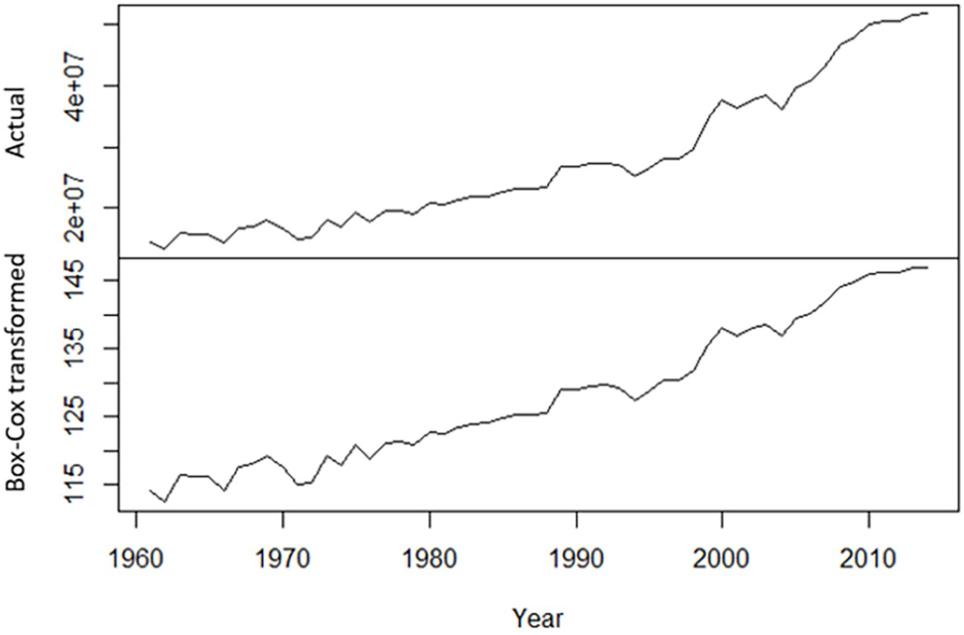

Fig 4. A comparison between the Box-Cox transformed sequence and the original sequence of annual rice production in Bangladesh.

https://doi.org/10.1371/journal.pone.0283452.g004

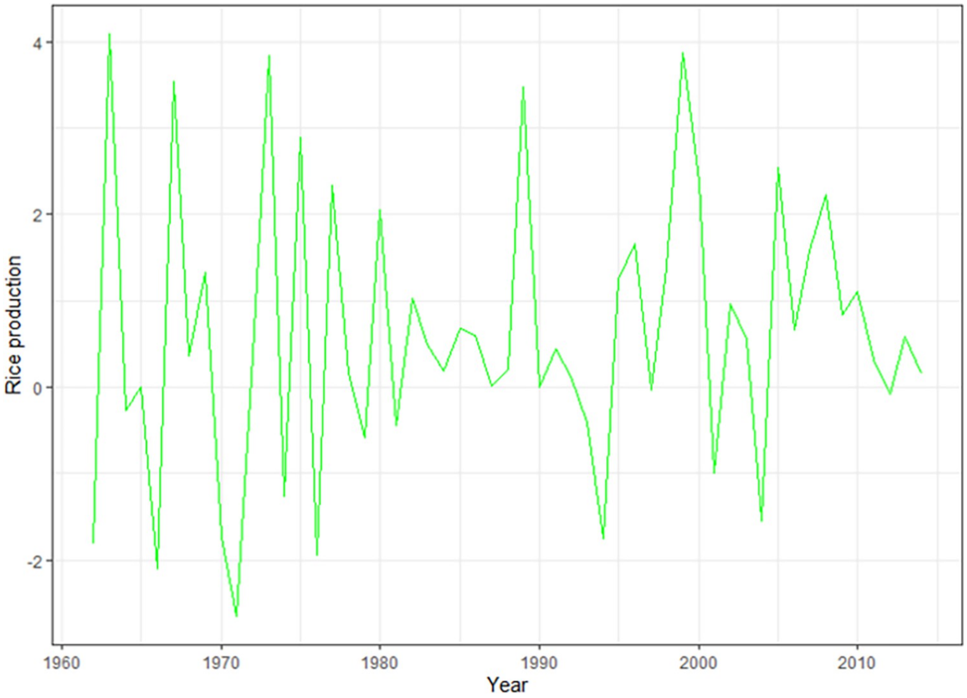

Fig 5. First-order differencing of the rice production of the training data set shows stationarity.

https://doi.org/10.1371/journal.pone.0283452.g005

We performed the ADF test to see the stationarity of the data and found the data non-stationary (p-value = 0.57) at level. To compensate for the trend shift observed in (Fig 4), we used first-order differencing of the transformed sequence (Fig 5). The differenced time series was found stationary using the ADF test (p-value = 0.01). So, the parameter (d) of the ARIMA model was 1.

In the ACF diagram, there was an evident peak at lag 1 indicating that the MA may become 1 and an evident spike at lags 0 in the PACF diagram, suggesting that the AR may become 0 (Fig 6). Therefore, the maximum p and q values are 0 and 1, respectively.

The ARIMA model was built with the "auto.arima" function to list all possible models and then selected the model ARIMA (0,1,1) with drift on the basis of the lowest Corrected Akaikes Information Criteria (AICc) value. The drift parameter value indicates that the rice production drifts upward positively (Table 1).

After that, the residual diagram, the ACF diagram of the residual, and the residual histogram were drawn, indicating a normal distribution (Fig 7). Hence, the ARIMA (0, 1,1) with drift model proved significant.

The XGBoost model was developed after adjusting several parameters. The adjusted parameters for the model were shown in S4 Table in S1 File. If a feature significantly affects the predicting performance when random noise takes its place, it is considered to be important. The feature importance of the XGBoost model was computed to see how each feature contributed to the prediction accuracy in the training set. And it was found that lag 5 of the training data contribute greatly to the model (Fig 8).

The curve of actual, fitted, and forecast values of the annual rice production in Bangladesh by ARIMA (0,1,1) with drift and the XGBoost model has been illustrated in Fig 9. The forecasted values of the XGBoost model were quite close to the actual values.

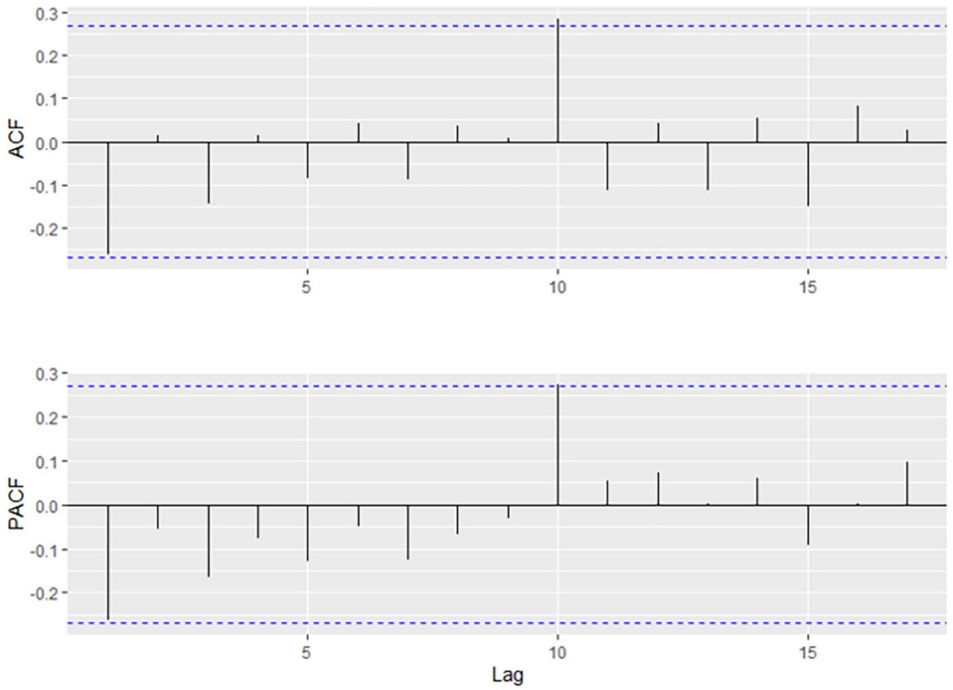

Fig 6. The ACF and PACF diagram of rice production in Bangladesh after first order differencing. ACF, autocorrelation function; PACF, partial autocorrelation function.

https://doi.org/10.1371/journal.pone.0283452.g006

Model comparison

The ARIMA (0,1,1) with drift model was built using the difference of the time series data. As a result, we lost a value in the training set; therefore, we compared the remaining 53 values. We used a maximum of eight time-lagged variables as input features for XGBoost. Because the maximum lag of 8 of the rice production data can contribute precisely to improve the XGBoost model prediction accuracy. Hence, the remaining 46 values were compared for the XGBoost model. The prediction accuracy for the ARIMA and XGBoost models is shown in Table 2.

The MAPE value of the test set of the XGBoost model was comparatively lower than the ARIMA model, which indicates that XGBoost performs better than ARIMA in predicting the annual rice production in Bangladesh. The detailed information regarding XGBoost model fitting can be found in S1 File.

Finally, based on our preferred XGBoost model, we predicted the annual rice production for the next 10 years (S1 File). According to our forecasts, during the next 10 years, the amount

Table 1. Estimated parameters of the ARIMA (0,1,1) with drift model.

Parameters	Estimate	Std. Error	z value	Pr(>|z|)
ma1	-0.32448	0.15445	-2.1008	0.03566*
drift	0.62942	0.14259	4.4142	0.00001***
AICc	201.54			

AICc: Corrected Akaikes Information Criteria

Std. Error: Standard Error

ARIMA: Autoregressive Integrated Moving Average

Asterisk (*) indicates significant at 1% and (***) indicates significant at 0% level.

https://doi.org/10.1371/journal.pone.0283452.t001

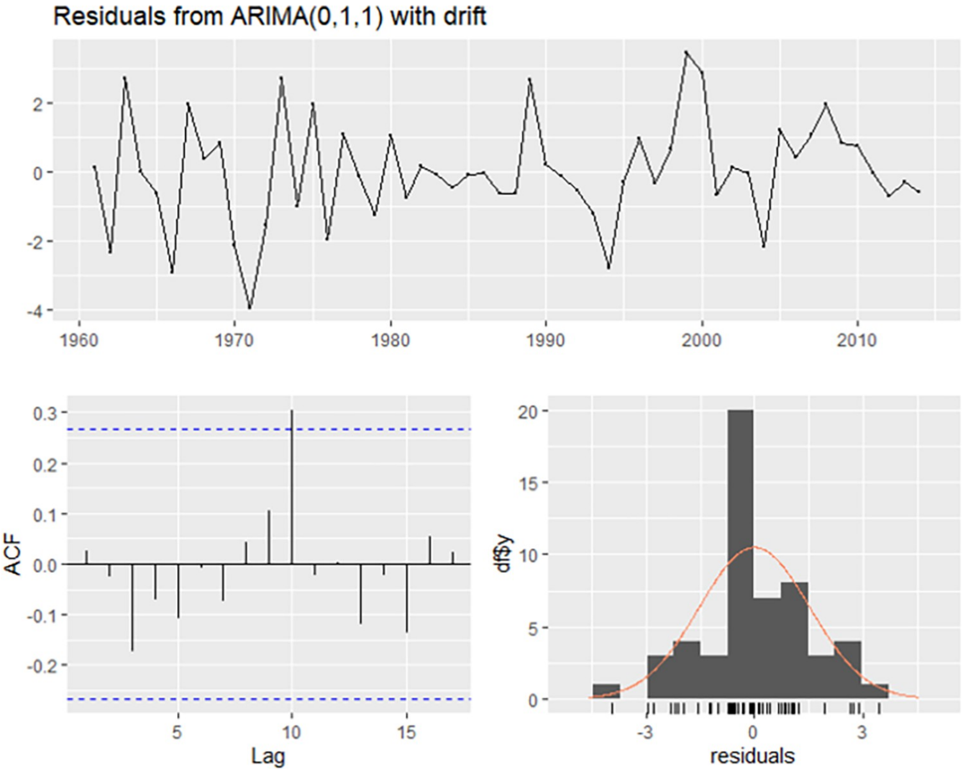

Fig 7. A time series plot of the residuals with corresponding ACF diagram, and a histogram for the ARIMA (0,1,1) model with drift. ACF, autocorrelation function; ARIMA, autoregressive integrated moving average.

https://doi.org/10.1371/journal.pone.0283452.g007

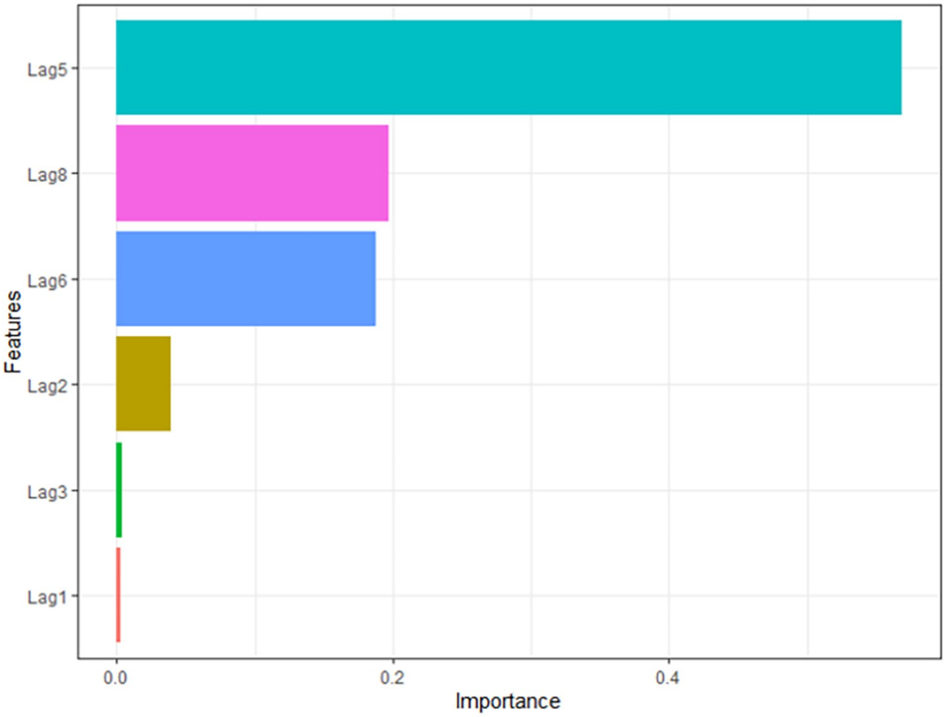

Fig 8. Important characteristic features of the XGBoost model.

https://doi.org/10.1371/journal.pone.0283452.g008

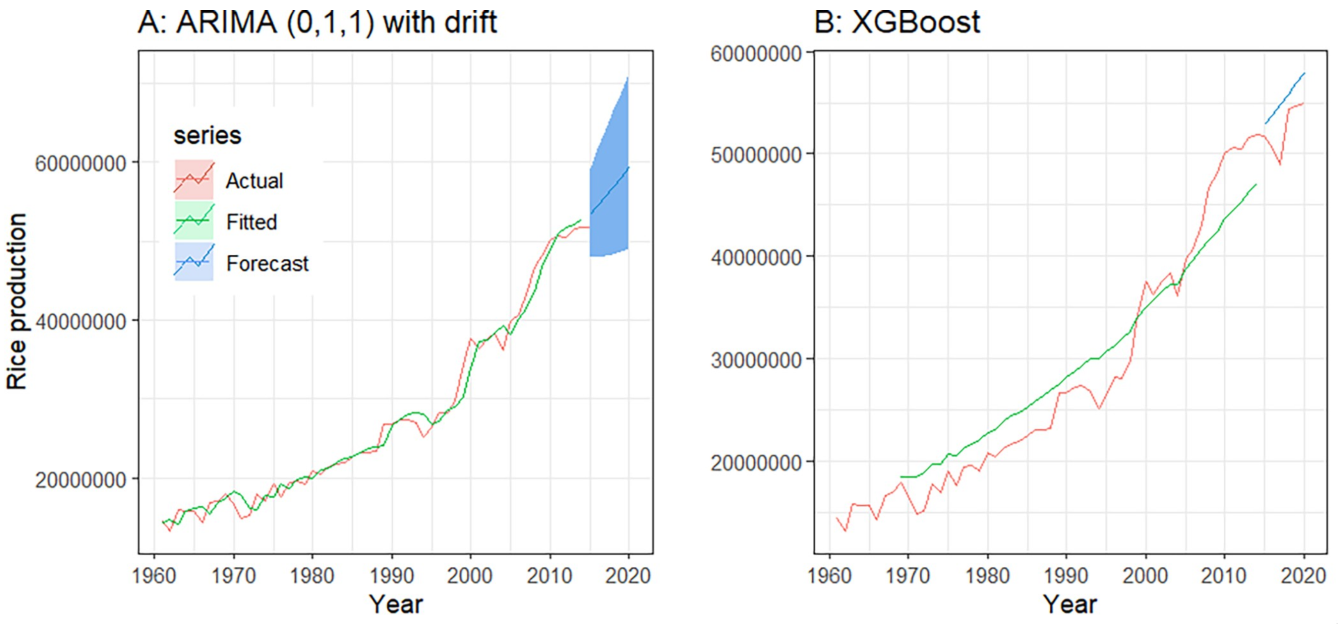

Fig 9. ARIMA and XGBoost model show the actual, fitted and forecasted data for rice production in Bangladesh. ARIMA, autoregressive integrated moving average; XGBoost, eXtreme Gradient Boosting.

https://doi.org/10.1371/journal.pone.0283452.g009

of rice produced annually in Bangladesh will vary between 57,850,318 and 82,256,940 tons, as illustrated in Fig 10.

Discussion

In our study, we found a linear upward pattern in the annual rice production data in Bangladesh. The primary goal of this study was to compare and contrast the predictive accuracy of the ARIMA and XGBoost forecasting models and make a short-term prediction with the best model. In this research, we examined the annual rice production in Bangladesh as a whole from 1961 to 2020. It is commonly known that Bangladesh has a subtropical tropical monsoon, which is distinguished by significant seasonal changes in precipitation, high temperatures, and humidity. In Bangladesh, there are three different seasons: a warm, humid summer from March to June; a chilly, wet monsoon season from June to October; and a cool, dry winter from October to March. In the past, temperatures in Bangladesh have ranged from 15˚C to

Table 2. Evaluation of parameters for the ARIMA and XGBoost model for rice production in Bangladesh.

Models	Training set				Test set			
	MAE	MPE	RMSE	MAPE	MAE	MPE	RMSE	MAPE
ARIMA(0,1,1)	1109886	-0.30	1496325	4.55	3755137	-7.23	4093961	7.23
XGBoost	2817876	-5.91	3209634	10.39	2779742	-5.39	3195985	5.38

ARIMA: Autoregressive Integrated Moving Average

MAE: Mean Absolute Error

MPE: Mean Percentage Error

MAPE: Mean Absolute Percentage Error

RMSE: Root Mean Square Error

XGBoost: eXtreme Gradient Boosting.

https://doi.org/10.1371/journal.pone.0283452.t002

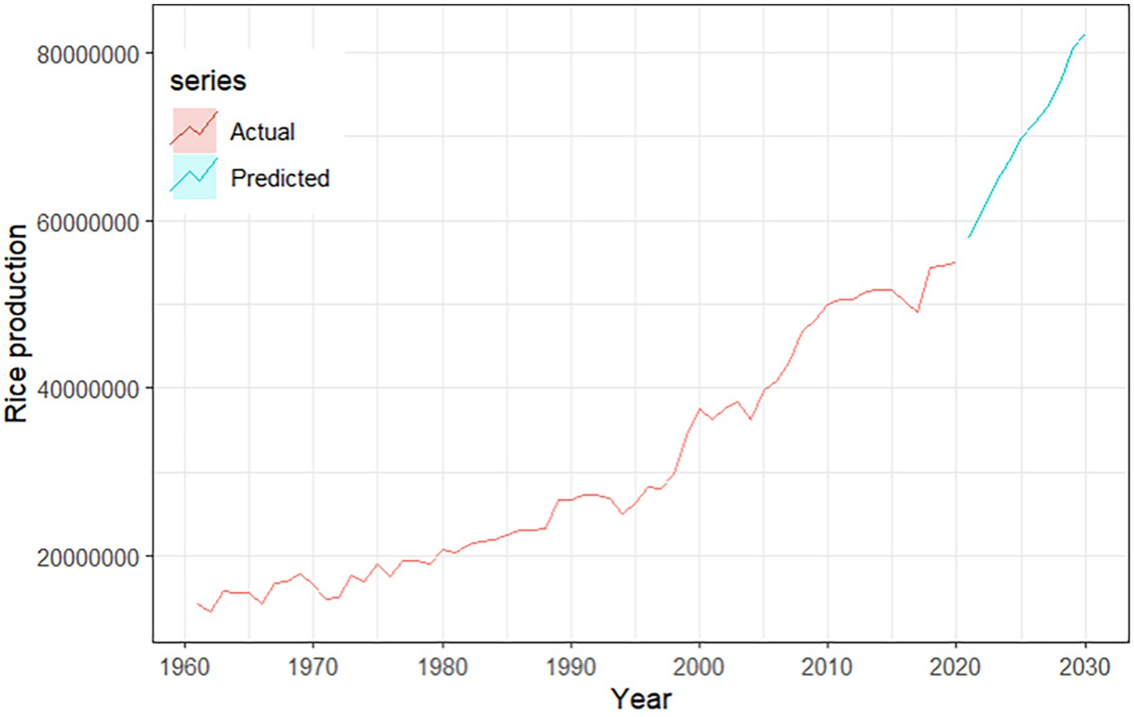

Fig 10. Ten years' prediction of annual rice production in Bangladesh using XGBoost model. XGBoost: eXtreme Gradient Boosting.

https://doi.org/10.1371/journal.pone.0283452.g010

34˚C annually, with an average temperature of roughly 26˚C [48, 49]. Food production (e.g. rice, wheat) is particularly vulnerable to climate change because the agricultural productions are severely impacted by the climate patterns. Several previous studies examined that mean temperature can negatively impact the rice production [50, 51]. Precipitation had a positive impact on rice production, which was also determined by a previous study [52]. To know the actual pattern of the annual rice production in Bangladesh and forecast it accurately, time series modeling is very crucial [53].

The ARIMA model for the annual rice production data was established based on the concept of linear regression to forecast future data points. Without using any other explanatory variable, the ARIMA model is capable of understanding the pattern of the historical data and making accurate forecasts. So, it is simple to establish the ARIMA model [24]. Since ARIMA is a well-known and most widely used time series forecasting model, this study compared the ARIMA model with the robust XGBoost machine learning model. The ARIMA model can be well fitted to non-stationary data after the Box-Cox transformation and differencing of the original data [39]. But differencing can cause data lose. In order to differencing the data, this study lost one-year data. We built the ARIMA models using the auto.arima function by adjusting the power transformation parameter (lambda) and selected the appropriate model based on the lowest AICc value. Based on the lowest AICc value, we finally selected the optimal ARIMA (0,1,1) with the drift model.

On the other hand, we used the tree-based ensemble XGBoost supervised machine learning technique on our data. Several previous studies used several machine learning models, such as the artificial neural network [22], the random forest [26, 54, 55], and the support vector machine [56, 57] to predict rice production and obtained effective predicting results. The eXtreme gradient boosting is a robust machine learning technique for precisely modeling, analyzing, and forecasting time series data [25]. The XGBoost model provides a variety of advantages regarding model forecasting. For example, it does not require any preprocessing of the data. It has a rapid processing speed, robust feature selection, good fitting, greater predictive performance and late scaling penalty than a typical Gradient boosting decision tree which removes the model from the occurrences of overfitting [25, 58]. As a result, we compared the predictive performance of the ARIMA model with the XGBoost model. From the result, it is clear that XGBoost performs better than the ARIMA model. In the meantime, the XGBoost model may also be utilized for cross-validation and has the ability to automatically identify significant feature vectors. The MAPE value of the XGBoost model for the test set is comparatively lower than the ARIMA model, which indicates XGBoost performs better than the ARIMA model in predicting the annual rice production in Bangladesh. Therefore, we used the XGBoost model to make a short-term prediction for the next 10 years. The prediction reveals that the amount of rice produced annually will grow in the following years in Bangladesh.

According to our study, the fitting and forecasting accuracy of the XGBoost model is much better than the traditional time-series ARIMA model. Without requiring any influencing factor, our proposed model can feasibly predict the annual rice production in Bangladesh.

Limitations

In this study, we identified a model by comparing the ARIMA and XGBoost models that could accurately predict the annual rice production in Bangladesh. There are several machine learning models such as Decision Tree, LightGBM, and so on that are more robust and might have greater prediction accuracy. These models need to be applied in the future to find the best one. We mainly concentrated on the effect of time on rice production, which made it simpler to develop and predict our model. As a result, one of the limitations is that some climatic and econometric factors like temperature, rainfall, consumption, and so on, which are well known to affect rice production, were not taken into account in this study. These should be investigated further in light of the data's availability.

Conclusion

We built an ARIMA and XGBoost model for forecasting the annual rice production in Bangladesh. These models were applied to generate a short-term prediction in this study. The XGBoost model performed better than the ARIMA model in predicting the annual rice production in Bangladesh. Finally, the government and development practitioners can employ XGBoost models over ARIMA to make more accurate short-term predictions of future crop production.

Supporting information

S1 File.
(DOCX)

Acknowledgments

We are very much grateful to the reviewers for providing valuable instructions and suggestions to make the study more appealing.

Author Contributions

Conceptualization: Mst Noorunnahar.

Data curation: Mst Noorunnahar, Arman Hossain Chowdhury, Farhana Arefeen Mila.

Formal analysis: Mst Noorunnahar, Arman Hossain Chowdhury.

Investigation: Mst Noorunnahar, Farhana Arefeen Mila.

Methodology: Mst Noorunnahar, Arman Hossain Chowdhury.

Resources: Mst Noorunnahar, Arman Hossain Chowdhury, Farhana Arefeen Mila.

Software: Arman Hossain Chowdhury.

Supervision: Mst Noorunnahar, Farhana Arefeen Mila.

Validation: Mst Noorunnahar, Arman Hossain Chowdhury, Farhana Arefeen Mila.

Visualization: Mst Noorunnahar, Arman Hossain Chowdhury, Farhana Arefeen Mila.

Writing – original draft: Mst Noorunnahar, Arman Hossain Chowdhury, Farhana Arefeen Mila.

Writing – review & editing: Mst Noorunnahar, Arman Hossain Chowdhury, Farhana Arefeen Mila.

References

1. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food Security: The Challenge of Feeding 9 Billion People. Science (80-). 2010; 327: 812–818. https://doi.org/10.1126/SCIENCE.1185383 PMID: 20110467
2. Rahman MC, Islam MA, Rahaman MS, Sarkar MAR, Ahmed R, Kabir MS. Identifying the Threshold Level of Flooding for Rice Production in Bangladesh: An Empirical Analysis. J Bangladesh Agric Univ. 2021; 19: 243–250. https://doi.org/10.5455/JBAU.53297
3. Khush GS. What it will take to Feed 5.0 Billion Rice consumers in 2030. Plant Mol Biol 2005 591. 2005; 59: 1–6. https://doi.org/10.1007/s11103-005-2159-5 PMID: 16217597
4. Dawe D. The contribution of rice research to poverty alleviation. Stud Plant Sci. 2000; 7: 3–12. https://doi.org/10.1016/S0928-3420(00)80003-8
5. Siddique MAB, Sarkar MAR, Rahman MC, Chowdhury A, Rahman MS, Deb L. Rice farmers' technical efficiency under abiotic stresses in Bangladesh. Asian J Agric Rural Dev. 2017; 7: 219–232. https://doi.org/10.18488/JOURNAL.1005/2017.7.11/1005.11.219.232
6. Sayeed KA, Yunus MM. Rice prices and growth, and poverty reduction in Bangladesh. 2018; 1–39. Available: http://www.fao.org/publications/card/en/c/I8332EN
7. BBS 2015. Statistical Yearbook of Bangladesh, Ministry of Planning, Government of the People's Republic of Bangladesh, Dhaka.
8. BBS 2020. Statistical Yearbook of Bangladesh, Ministry of Planning, Government of the People's Republic of Bangladesh, Dhaka.
9. Bangladesh Economic Review 2020. Economic Adviser's Wing, Finance Division, Ministry of Finance, Government of the People's Republic of Bangladesh.
10. Gebbers R, Adamchuk VI. Precision Agriculture and Food Security. Science (80-). 2010; 327: 828–831. https://doi.org/10.1126/science.1183899 PMID: 20150492
11. Ji Z, Pan Y, Zhu X, Wang J, Li Q. Prediction of Crop Yield Using Phenological Information Extracted from Remote Sensing Vegetation Index. Sensors 2021, Vol 21, Page 1406. 2021; 21: 1406. https://doi.org/10.3390/s21041406 PMID: 33671356
12. Kumar N. A Novel Method for Rice Production Forecasting Using Fuzzy Time Series. Int J Comput Sci Issues. 2012; 9: 455–459.
13. Alam W, Mrinmoy RAY, Kumar RR, Sinha K, Rathod S, Singh KN. Improved ARIMAX modal based on ANN and SVM approaches for forecasting rice yield using weather variables. Indian J Agric Sci. 2018; 88: 1909–1913.
14. Jing-feng HU, Zhong-en YA, Ren-chao WA, Hong-wei XU HJ. The rice production forecasting models using NOAA/AVHRR data based on GIS. Remote Sens Technol Appl. 2011; 17: 125–128.
15. Yun JI. Predicting regional rice production in South Korea using spatial data and crop-growth modeling. Agric Syst. 2003; 77: 23–38. https://doi.org/10.1016/S0308-521X(02)00084-7

16. Koide N, Robertson AW, Ines AVM, Qian JH, Dewitt DG, Lucero A. Prediction of rice production in the Philippines using seasonal climate forecasts. J Appl Meteorol Climatol. 2013; 52: 552–569. https://doi.org/10.1175/JAMC-D-11-0254.1

17. Noureldin NA, Aboelghar MA, Saudy HS, Ali AM. Rice yield forecasting models using satellite imagery in Egypt. Egypt J Remote Sens Sp Sci. 2013; 16: 125–131. https://doi.org/10.1016/j.ejrs.2013.04.005

18. Bandumula N. Rice Production in Asia: Key to Global Food Security. Proc Natl Acad Sci India Sect B Biol Sci 2017 884. 2017; 88: 1323–1328. https://doi.org/10.1007/S40011-017-0867-7

19. Rahman NMF, Hasan MM, Hossain MI, Baten MA, Hosen S, Ali MA, et al. Forecasting Aus Rice Area and Production in Bangladesh using Box-Jenkins Approach. Bangladesh Rice J. 2016; 20: 1–10. https://doi.org/10.3329/BRJ.V20I1.30623

20. Mahmud S. Predicting the Rice Production of Bangladesh by Machine Learning Technique. 2018; 7: 7–13.

21. Rahman N. Forecasting of boro rice production in Bangladesh: An ARIMA approach. J Bangladesh Agric Univ. 1970; 8: 103–112. https://doi.org/10.3329/JBAU.V8I1.6406

22. Sultana A, Khanam M. Forecasting Rice Production of Bangladesh Using ARIMA and Artificial Neural Network Models. Dhaka Univ J Sci. 2020; 68: 143–147. https://doi.org/10.3329/DUJS.V68I2.54612

23. Rodríguez JP, Corrales DC, Griol D, Callejas Z, Corrales JC. A Non-Destructive Time Series Model for the Estimation of Cherry Coffee Production. C Mater Contin. 2022; 70: 4725–4743. https://doi.org/10.32604/CMC.2022.019135

24. Lv CX, An SY, Qiao BJ, Wu W. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. BMC Infect Dis. 2021; 21: 1–13. https://doi.org/10.1186/S12879-021-06503-Y/TABLES/5

25. Alim M, Ye GH, Guan P, Huang DS, Zhou B Sen, Wu W. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: A time-series study. BMJ Open. 2020; 10: 1–8. https://doi.org/10.1136/bmjopen-2020-039676 PMID: 33293308

26. Narasimhamurthy V. Rice Crop Yield Forecasting Using Random Forest Algorithm SML. Int J Res Appl Sci Eng Technol. 2017; V: 1220–1225. https://doi.org/10.22214/ijraset.2017.10176

27. Anitha P, Chakravarthy T. Agricultural Crop Yield Prediction using Artificial Neural Network with Feed Forward Algorithm. Int J Comput Sci Eng. 2018; 6: 178–181. https://doi.org/10.26438/ijcse/v6i11.178181

28. Aler R, Galván IM, Ruiz-Arias JA, Gueymard CA. Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. Sol Energy. 2017; 150: 558–569. https://doi.org/10.1016/J.SOLENER.2017.05.018

29. Fang ZG, Yang SQ, Lv CX, An SY, Wu W. Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series study. BMJ Open. 2022; 12: 1–8. https://doi.org/10.1136/bmjopen-2021-056685 PMID: 35777884

30. Rahman MS, Chowdhury AH. A data-driven eXtreme gradient boosting machine learning model to predict COVID-19 transmission with meteorological drivers. 2022; 1–14. https://doi.org/10.1371/journal.pone.0273319 PMID: 36099253

31. Khashei M, Bijari M, Raissi Ardali GA. Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs). Comput Ind Eng. 2012; 63: 37–45. https://doi.org/10.1016/J.CIE.2012.01.017

32. Pai PF, Lin CS. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega. 2005; 33: 497–505. https://doi.org/10.1016/J.OMEGA.2004.07.024

33. Kabir MS, Salam MU, Chowdhury A, Rahman MF, Iftekharuddaula KM, Rahman MS, et al. Rice Vision for Bangladesh: 2050 and Beyond. Bangladesh Rice J. 2015; 19: 1–18. https://doi.org/10.3329/BRJ.V19I2.28160

34. FAOSTAT. Annaul Rice Production data of Bangladesh. [cited 8 Dec 2022]. Available: https://www.fao.org/faostat/en/#data

35. Helfenstein U. Box-Jenkins modelling in medical research. 2016; 5: 3–22. https://doi.org/10.1177/096228029600500102 PMID: 8743076

36. Amin M, Amanullah M, Akbar A. Time series modeling for forecasting wheat production of Pakistan. J Anim Plant Sci. 2014; 24: 1444–1451.

37. Alzahrani SI, Aljamaan IA, Al-Fakih EA. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health. 2020; 13: 914–919. https://doi.org/10.1016/j.jiph.2020.06.001 PMID: 32546438

38. Sahai AK, Rath N, Sood V, Singh MP. ARIMA modelling & forecasting of COVID-19 in top five affected countries. Diabetes Metab Syndr Clin Res Rev. 2020; 14: 1419–1427. https://doi.org/10.1016/J.DSX.2020.07.042 PMID: 32755845

39. Rahman MS, Chowdhury AH, Amrin M. Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. Plos Glob Public Heal. 2022; 2019: 1–13. https://doi.org/10.1371/journal.pgph.0000495

40. Li W, Yin Y, Quan X, Zhang H. Gene Expression Value Prediction Based on XGBoost Algorithm. Front Genet. 2019; 10: 1–7. https://doi.org/10.3389/fgene.2019.01077 PMID: 31781160

41. Luo J, Zhang Z, Fu Y, Rao F. Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. Results Phys. 2021; 27: 104462. https://doi.org/10.1016/j.rinp.2021.104462 PMID: 34178594

42. Paliari I, Karanikola A, Kotsiantis S. A comparison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting. IISA 2021 - 12th Int Conf Information, Intell Syst Appl. 2021. https://doi.org/10.1109/IISA52424.2021.9555520

43. Prajapati S, Swaraj A, Lalwani R, Narwal A, Verma K, Singh G. Comparison of Traditional and Hybrid Time Series Models for Forecasting COVID-19 Cases. 2019;8.

44. RStudio: Integrated Development Environment for R RStudio Team. In: RStudio, PBC, Boston, MA (2022) [Internet]. [cited 18 Dec 2022]. Available: https://www.rstudio.com/

45. Hyndman RJ, Khandakar Y. Automatic Time Series Forecasting: The forecast Package for R. J Stat Softw. 2008; 27: 1–22. https://doi.org/10.18637/JSS.V027.I03

46. Sakia RM. The Box-Cox Transformation Technique: A Review. Stat. 1992; 41: 169. https://doi.org/10.2307/2348250

47. Curran-Everett D. Explorations in statistics: The log transformation. Adv Physiol Educ. 2018; 42: 343–347. https://doi.org/10.1152/advan.00018.2018 PMID: 29761718

48. Bangladesh - Climatology | Climate Change Knowledge Portal. [cited 13 Dec 2022]. Available: https://climateknowledgeportal.worldbank.org/country/bangladesh/climate-data-historical

49. Climate of the World: Bangladesh | weatheronline.co.uk. [cited 18 Dec 2022]. Available: https://www.weatheronline.co.uk/reports/climate/Bangladesh.htm

50. Stuecker MF, Tigchelaar M, Kantar MB. Climate variability impacts on rice production in the Philippines. PLoS One. 2018;13. https://doi.org/10.1371/journal.pone.0201426 PMID: 30091991

51. Pickson RB, He G, Boateng E. Impacts of climate change on rice production: evidence from 30 Chinese provinces. Environ Dev Sustain 2021 243. 2021; 24: 3907–3925. https://doi.org/10.1007/S10668-021-01594-8

52. Mahmood N, Ahmad B, Hassan S, Bakhsh K. Impact of temperature ADN precipitation on rice productivity in rice-wheat cropping system of Punjab province. J Anim Plant Sci. 2012; 22: 993–997.

53. Reddy PCS, Sureshbabu A. An Applied Time Series Forecasting Model for Yield Prediction of Agricultural Crop. Adv Intell Syst Comput. 2020; 1118: 177–187. https://doi.org/10.1007/978-981-15-2475-2_16/COVER/

54. Kim J, Lee J, Sang W, Shin P, Cho H, Seo M. Random Forest를 이용한 남한지역 쌀 수량 예측 연구 Rice yield prediction in South Korea by using random forest. 2019; 21: 75–84. https://doi.org/10.5532/KJAFM.2019.21.2.75

55. Choudhary K, Shi W, Dong Y, Paringer R. Random Forest for rice yield mapping and prediction using Sentinel-2 data with Google Earth Engine. Adv Sp Res. 2022; 70: 2443–2457. https://doi.org/10.1016/J.ASR.2022.06.073

56. Fegade TK, Pawar B V. Crop Prediction Using Artificial Neural Network and Support Vector Machine. Adv Intell Syst Comput. 2020; 1016: 311–324. https://doi.org/10.1007/978-981-13-9364-8_23/COVER

57. Gandhi N, Petkar O, Armstrong LJ, Tripathy AK. Rice crop yield prediction in India using support vector machines. 2016 13th Int Jt Conf Comput Sci Softw Eng JCSSE 2016. 2016. https://doi.org/10.1109/JCSSE.2016.7748856

58. Wu W, Guo J, An S, Guan P, Ren Y, Xia L, et al. Comparison of two hybrid models for forecasting the incidence of hemorrhagic fever with renal syndrome in Jiangsu Province, China. PLoS One. 2015; 10: 1–13. https://doi.org/10.1371/journal.pone.0135492 PMID: 26270814