Forecasting Tea Production in the Context of Bangladesh Utilizing Machine Learning

Preprint · July 2023

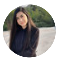
DOI: 10.13140/RG.2.2.30536.70408

5 authors, including:


Abdullah Al Ryan
Daffodil International University


Silvia Kh Shuvessa
Daffodil International University


Shahriar Mamun


Habiba Dewan Arpita
Daffodil International University

Forecasting Tea Production in the Context of Bangladesh Utilizing Machine Learning

Abdullah Al Ryan
dept. of CSE
Daffodil International University
Dhaka, Bangladesh
abdullah15-13088@diu.edu.bd

Silvia Kh Shuvessa
dept. of CSE
Daffodil International University
Dhaka, Bangladesh
silvia15-12261@diu.edu.bd

Shahriar Mamun
dept. of CSE
Daffodil International University
Dhaka, Bangladesh
shahriar15-12859@diu.edu.bd

Habiba Dewan Arpita
dept. of CSE
Daffodil International University
Dhaka, Bangladesh
habiba15-14042@diu.edu.bd

Mr. Md. Sazzadur Ahamed
dept. of CSE
Daffodil International University
Dhaka, Bangladesh
sazzad.cse@diu.edu.bd

***Abstract*— One of the most popular drinks, second only to water, is tea. Bangladesh ranks as the world's 10th-largest manufacturer of tea. Tea has a big impact on poverty reduction, rural development, and nutrition security. Over the past ten years, Bangladesh has increased its tea supply. According to the BTB figures, over 96.51 million kg were produced in 2021, an increase of almost 54% from that of 2012. As tea is a profitable crop from Bangladesh's perspective, tea yield prediction can play a significant role in increasing the production of tea. In this paper, tea yield has been forecast using various machine learning algorithms based on area-based tea production and their weather data from 1968 to 2021. Necessary data has been collected from BBS, BARC, and BMD government organizations. Eight climate factors have been used for the research. The data model has been evaluated by using eight classification and regression algorithms, with the Random Forest classifier showing the best accuracy of 97%. With this approach, tea production can be increased while reducing food threats and managing it for other nations.**

Keywords*— *Yield, Tea, Prediction, Machine Learning, Production, RFR, RFC, Bangladesh, DTR, GBR, AdaBoost, Forecast, Accuracy.

I. INTRODUCTION

In Bangladesh, agriculture is practiced over more than half of the country's total land area of 15 million hectares [1]. Modern agriculture must produce more food to meet the demands of a fast-expanding global population. As a result, the agricultural industry is utilizing the most recent technology to increase net production through the collection and processing of information [2]. 11.63 percent of Bangladesh's GDP in 2021 was attributable to agriculture [3]. Tea, the 2nd largest cash crop, has long been a favorite beverage, enjoyed from breakfast to late-night conversation. As a result, Bangladesh has produced tea for more than 180 years. Bangladesh presently has 167 commercial tea-producing plantations and gardens in operation on 2,79,507.88 acres of land, employing over 150K people. Furthermore, Bangladesh generates 3% of global tea production. The Bangladeshi tea market was estimated to be worth BDT 3500 crore in 2021. Bangladesh produced the most tea ever-9 crores, 65 lac kg in 2021. In the same year, Bangladesh earned $180.57 million by exporting 680,000 kg of tea worldwide. [5].

Applying automated learning and scientific methodologies to solve problems has evolved from a fad to a requirement in our technological age. The first-world countries have improved in areas like medicine, agriculture, education, and other disciplines as a result of this knowledge. Finding hidden patterns in massive datasets and connecting them to solve an issue through data analysis is known as "data mining." Data mining's introduction to the agriculture sector has benefited research. A significant issue in agriculture is yield prediction. Every farmer wants to know how much of a harvest to anticipate. In the past, farmer expertise with a particular crop was taken into account when predicting production [6]. The tea business has a huge amount of data.

It is challenging to anticipate the yearly tea production due to climate change, global warming, and other natural disasters. Bangladesh's diminishing arable lands, floods, salt intrusion, and drought have traditionally endangered the industry's survival. Because the weather in different areas of Bangladesh varies, it's crucial to take into account environmental characteristics unique to each of these regions. This will make it easier to select the optimum areas for the development of various crops. Additionally, rainfall varies greatly from area to area and district. While less or huge rainfall might harm crops, the perfect quantity of rain results in the best agricultural output [7]. Because rainfall as well as humidity, which is a side effect of rainfall, changes from region to region, To achieve the right yield, a region consisting of the ideal average annual humidity and rainfall is needed. Humidity alters the amount of water that the atmosphere can absorb, which might keep crops too wet or too dry [8]. Data on cloud cover, wind speed, and sunshine have also been gathered for this study. These variables differ from location to location. Temperature and moisture both affect cloud cover.

The techniques of machine learning covered in this work are RFC, RFR, DTC, DTR, GBR, LR, KNN, and AdaBoost. They are applied to all data from 23 regions of Bangladesh. Meteorological data, amount of tea production (per ton), and cultivated area (per acre) are collected from BMD, BBS, and BARC. Our contributions are as follows:

- A new custom dataset with 11 attributes has been created which are collected from government organizations consisting of all the regions of Bangladesh from the year 1968 to 2021 across 54 years.
- This study on Bangladeshi Tea yield with real-time data has not yet been conducted.
- Ensured that the algorithms obtained remarkable precisions. Specifically, the Random Forest classifier has provided the best accuracy with 97% on our real-time datasets.

By applying machine learning approaches that can help in the prediction of future trends and productivity, this study aims to close the knowledge gap. The structure of this article is as follows: Division II provides a concise summary of the earlier works. The experimental setup and suggested method for predicting rice production are detailed in Section III. Section IV includes a description of the assessment procedures and their outcomes. Section V comes to an end with the article.

II. Related Work

Batool et al.[9] in their research, compared ten different regression models and the Aqua Crop simulation model using agro-management, crop, weather and soil data from the tea fields of Pakistan's NTHRI from 2016 to 2019. Among all other regression algorithms, XGBoost regression exceeded other ML regression methods and achieved a RMSE of 0.154.Palanivel and Suryanarayana [10] came up with a conceptual idea for implementing big data computing in machine learning approaches to predict crop yield.

Burhan [11] used annual yield data from 1990 to 2019 to use regression algorithms like DTR, RFR, and SVR algorithms to make predictions about nine important crops in Turkey: barley, grapes, apples, maize, olives, sugar, tomatoes, potatoes, wheat, and beets. The study found that DTR and RFR can accurately predict the yields of wheat, barley, and maize with 94% and 87% accuracy, respectively. On the other hand, SVR results in an unstable forecast.Ahmed et al. [12] applied SVR, GPR, ANN, and RF algorithms. Ann had the best R2 value of 0.9461 and a RMSE of 0.1204, but it wasn't stable, so neural networks were ensembled to reduce the variation in the ANN model. collected every month and obtained from 2008 to 2018, for a total of 120 samples. There are nine independent variables in the data: the rainfall, the minimum and maximum temperature of the farm, the humidity level, and the PH value of the soil. Jambekar et al. [13]compiled data from the Indian government's publicly available records from 1950 to 2013 and applied ML algorithms; they discovered that MARS (Earth) performed better than MLR and RFR on the rice and datasets, and MLR performed better than RFR and MARS (Earth) on the maize dataset Silas and Nderu [14] did Kenya-based research using the clustering and association rule data mining approaches to speculate on tea production in Kenya.Baruah et al. [15] introduced a strategy to envision the forthcoming crop productivity from four tea-growing regions in Assam: Cachar, the NoBank, Uth Bank, and Upper Assam. The technique incorporated the MLR technique, and corresponding data were collected from four meteorological stations.

Das et al. [16] This paper is Bangladeshi. The authors analyzed GIS applications using satellite remote sensing, and the analytical hierarchy placed 3.37% of the land in the highly suitable category. In Sylhet Division, 9.01% of the land was moderately suitable, 49.87% was marginally suitable, and 17.1% was unsuitable for tea cultivation.Duncan et al. [17] derived data between 2004 and 2013 from 82 tea gardens in Assam. Statistical models were applied to ascertain how monthly temperature, precipitation, the severity of the current drought, and variance in monthly rainfall affect tea yield. Meteorological and NDVI data were integrated by Phan et al. [18] to produce high-yield forecast performance. SVM, RF, and TLRM were three ML models employed to forecast tea yield. The RF model's forecast tea yield for the years 2009 to 2018 had the highest R2 value of 0.73, followed by the SVM 0.66 and the TLRM's 0.57.A composite DRS-RF spatiotemporal model approach with an optimization technique and SVR to select features was introduced by Jui et al. [19]. When the suggested method was compared to various existing non-hybrid algorithms, it produced results that were more satisfying. Islam et al. [20] used secondary data from 1990 to 2015 to anticipate domestic tea consumption and manufacture in Bangladesh by employing an auto-regressive integrated moving average model for the following five years.

Five Gandhi et al. [21] investigated factors such as reference crop evapotranspiration, precipitation area, average temperature, lowest temperature, production, highest temperature, and yield throughout their analyses of 27 Maharashtra districts between 1998 and 2002, from June to November. Researchers assessed the classification algorithms BayesNet and NaiveBayes to predict how much rice will grow in Maharashtra, India, during the Kharif season. They found that BayesNet outperformed Naive Bayes in terms of performance. By combining machine learning techniques with satellite images and meteorological data, Schwalbert et al. [22] predicted soybean output in southern Brazil. A more accurate yield projection was made possible by the merging of meteorological information and satellite images. Hammer et al. [23] compared the DM techniques and an agroecological zone simulation model like GBM, RF, and SVM for forecasting sugarcane yield and came to the conclusion that DM approaches are more satisfactory than an agroecological model. Khaki and Wang [24] achieved an RMSE of 11% by distinguishing corn hybrids and genotype or environment data and using the ML approach to predict them accurately. Yethiraj [25] analyzed that data mining techniques like SSVM, NN, KNN, K-means, ID3, and association rule mining could be used to find the same sequence in the data and support production forecasting in the future of the agricultural sector.

TABLE I. SOME CURRENT, UNIQUE AGRICULTURAL YIELD RESEARCH IN MACHINE LEARNING

Citation	Year	Algorithm	Data	Accuracy
Batool et al. [9]	2022	Ten different regression models	2016-2019	XGBoost had the best accuracy • MAE:0.123 t/ha •MSE:0.024 t/ha • RMSE:0.154t/ha
Phan et al. [18]	2020	SVM, RF, and TLRM were applied to forecast tea yield	2009-2018	RF gave the best accuracy with R2 value: 0.73
Silas and Nderu [14]	2017	Clustering and Association Rule Mining	2003-2015	Summarized average rate of tea production in variables: • Extremely high=0.23 • High=2.9 •Average=5.38 • Below Average=3.15 • Extremely Low= 0.0769

Baruah et al. [15]	2016	Multiple Linear Regression (MLR)	1977-2006	• For South Bank, R=82 • North Bank, R=77% • Cachar, R=65%
Our work	2023	• LR • GBR • DT • RF • KNN • AdaBoost	1968-2018	RF showed the best accuracy with 97%

III. METHODOLOGY

The complete study procedure is delimited into five parts. Figure 1 depicts the complete working procedure, which is thoroughly discussed in the parts that follow.

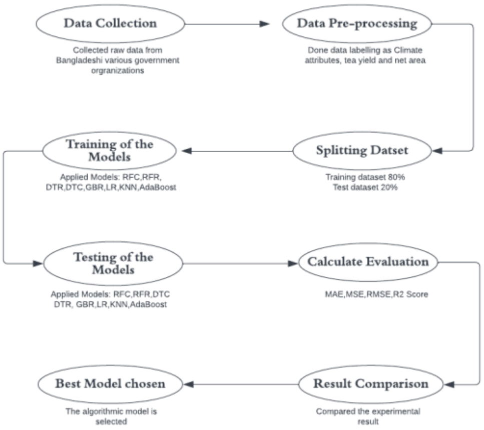


Figure 1. Flow Diagram of ML Models

A. Data Collection and Data Analysis

The raw data was obtained from the Bangladesh Bureau of Statistics (BBS). The yearly books of the BBS were used to gain information on the tea production (bales) and area (ha) in the targeted region, and their weather database was used to obtain the data on the weather. The meteorological variables that affect tea production the most are rainfall, maximum and minimum temperatures, cloud cover, humidity, sunshine, and wind speed. Consequently, yearly data on these climatic factors were gathered.

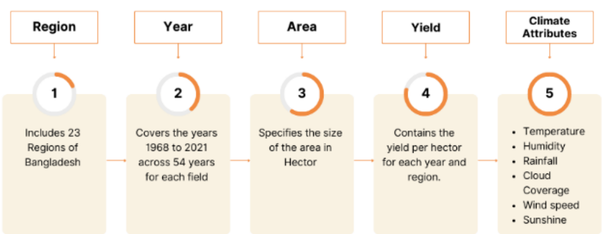
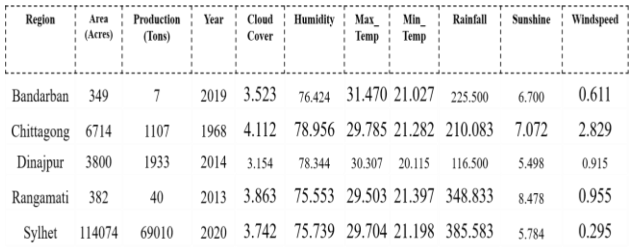


Figure 2: Tea Yield Dataset



Region	Area (Acres)	Production (Tons)	Year	Cloud Cover	Humidity	Max_ Temp	Min_ Temp	Rainfall	Sunshine	Windspeed
Bandarban	349	7	2019	3.523	76.424	31.470	21.027	225.500	6.700	0.611
Chittagong	6714	1107	1968	4.112	78.956	29.785	21.282	210.083	7.072	2.829
Dinajpur	3800	1933	2014	3.154	78.344	30.307	20.115	116.500	5.498	0.915
Rangamati	382	40	2013	3.863	75.553	29.503	21.397	348.833	8.478	0.955
Sylhet	114074	69010	2020	3.742	75.739	29.704	21.198	385.583	5.784	0.295

Figure 3. Sample of Dataset

B. Data Preprocessing

Processing the data was done both before and after the dataset was created. Prior to starting our data, we separated the entire data set according to the region to get region-specific information. The weather information was still updated monthly. To obtain the yearly climate data, we average the climate data from the previous 12 months. After creating the dataset, our goal was to model construction. The null value was initially eliminated. The categorical data is then preprocessed using a label encoder. To feed the machine, we transform categorical region data into numeric data.

C. Data Exploration

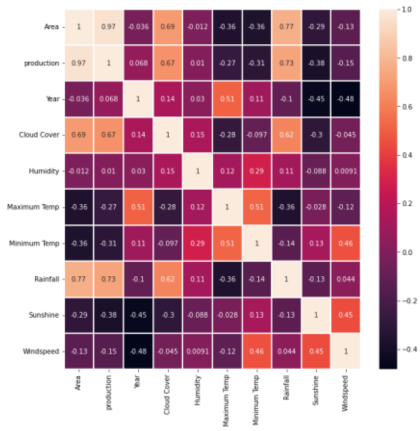


Figure 4. Tea Bangla Dataset

A correlation matrix is a matrix that shows the correlation coefficients between several variables. Each cell of the matrix has a depiction of the correlation between two variables. Correlation matrices are used to summarize data, and they are also used as inputs and diagnostics for more complex investigations and analyses.

D. Data Segmentation

To avoid overfitting, machine learning often utilizes data segmentation. In this case, a machine learning model cannot consistently match the starting data because it matches its training data too well. The datasets used in this study were divided into two categories. We utilized 20% for testing and 80% for training the model, respectively.

E. Machine Learning Models

1) Linear Regression: Linear regression is a basic and often employed machine learning technique. This method is used to conduct predictive analysis. Predictions are made using linear regression for continuous real or numeric variables like temperature, humidity, rainfall, and productivity, among others. Multiple regression, one of two types of linear regression, was used in this instance.

2) Gradient Boosting Regressor: Classification and regression issues can be solved using the machine learning algorithm gradient boosting. It provides a decision tree regression prediction model, which is comparable to inadequate estimating methods. Gradient-boosted trees, the alternative approach used when a decision tree is a bad learner, frequently outperforms random forests. A gradient-boosted tree model is constructed similarly to earlier boosting methods, stage by stage, but it generalizes prior methods by enabling the optimization of any discrete weight vector.

3) Decision Tree: We may use decision tree approaches for both classification and regression problems, which is different from other supervised learning techniques. The goal is to utilize a decision tree to learn fundamental decision rules from time series data in order to create a training model that can be used to forecast the class or value of the target attribute (training data).

4) Random Forest: A sizable fraction of the decision trees employed by the RF method, which is used for classification, regression, and other issues, are constructed during the training stage. The result of the RFt is the class to which the majority of the trees in a classification task were allocated. For regression tasks, the average prediction generated by a certain tree is supplied. Random choice forests are appropriate for this since decision trees typically overfit their training set. Even though they usually beat decision trees, RFs are less accurate than gradient-enhanced trees. However, the quality of the data may affect how successful they are.

5) KNN: KNN is an easy-to-use machine learning technique that is founded on the supervised learning methodology. On the premise that the new and existing examples are comparable, the KNN approach places the new instance in the category that closely resembles the existing categories. After storing all the previous data, a new data point is classified using the K-NN algorithm based on similarity. This shows that utilizing the KNN approach, fresh data may be consistently and efficiently classified. Although the KNN technique is most usually used to solve classification issues, it may also be used to solve regression difficulties. K-NN is a non-parametric approach, hence no assumptions are made about the underlying data.

6) Ada Boost Regressor: Machine learning uses the adaptive boosting methodology, sometimes known as the AdaBoost algorithm, as part of its ensemble approach. The weights are reallocated to each instance, with greater weights being given to instances that were incorrectly classified—hence, the term "adaptive boosting"—and each instance is given a new set of weights.

F. Error and Score Models

1) Mean Absolute Error (MAE): The Mean Absolute Error is a measure of the average absolute difference between the actual and expected values in the dataset. It determines the average residuals for the dataset.

2) Mean Squared Error (MSE): The Mean Squared Error is the average of the squared difference between the original and predicted values of the data set. It determines the variance of the residuals.

3) Root Mean Squared Error (RMSE): Root Mean Squared Error is the name for Mean Squared Error's square root. The standard deviation of the residuals is calculated.

4) R2 Score: The coefficient of determination, commonly known as R-squared, shows how much of the variance in the dependent variable the linear regression model can explain. The score is scale-free, so it will always be less than one regardless of how big or small the numbers are.

IV. RESULT AND DISCUSSION

Figure 5 shows how the Random Forest classifier surpasses the other methods in terms of accuracy. The highest accuracy was provided by the Random Forest Classifier at 97%. Both the Gradient Boosting and the Forest Regressors had an accuracy of 96%. Ada boosting regressor accuracy was 93%, and linear regression and KNN accuracy were both 88%. The decision tree classifier and regressor have the lowest accuracy rates for this data set at 87% and 82%, respectively.

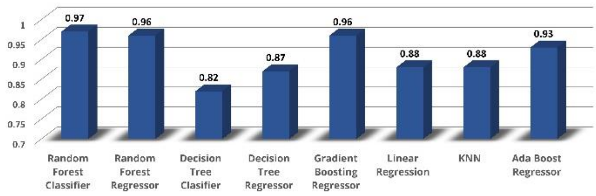


Figure 5: Accuracy Score

The eight machine learning-based models (LR, GBR, RFR, RFC, DTC, DTR, KNN, and AdaBoost) for estimating tea yield all perform similarly, as shown in Figure 6. Despite the models' comparable performance, the one created using the Random Forest Classifier approach came out marginally superior, with the highest r2 score and the lowest MAE and RMSE.

ALGORITHMS	MAE	MSE	RMSE	R2 SCORE
Random Forest Classifier	886.542	6258226.142	2501.644	0.97
Random Forest Regressor	904.021	9717624.412	3117.31	0.96
Decision Tree Classifier	2228.42	43894100.37	6625.26	0.82
Decision Tree Regressor	1378.34	32120613.42	5667.50	0.874
Gradient Boosting Regressor	852.63	8629683.60	2937.63	0.96
Linear Regression	2601.84	29837776.13	5462.39	0.88
KNN	2100.4	29633676.0	5443.6	0.88
Ada Boost Regressor	1299.002	17134275.301	4139.356	0.93

Figure 6: Error Score

superior, with the highest r2 score and the lowest MAE and RMSE.

V. CONCLUSION AND FUTURE WORK

Using One of Bangladesh's primary exports is tea. The age of the tea plants, variations in temperature, illnesses, and precipitation have all contributed to a stagnant tea output in recent years. Tea status monitoring and yield forecasting are required in this situation. However, there is little research on tea. We have introduced a machine-learning strategy for forecasting tea production. The study's overall results may be summed up as follows:

- A unique dataset consisting of the information on tea yield which was collected from BBS of Bangladesh from 1967 to 2021 has been created.
- By far the best performance in forecasting tea yield has been demonstrated by the suggested RFC model, which has an accuracy of 97%, while RFC, RFR, DTC, DTR, GBR, LR, KNN, and AdaBoost Regressor have an accuracy of 97%, 96%, 82%, 87%, 88%, and 93%, respectively.
- The Random Forest Classifier method had the best R2 score and the lowest MAE and RMSE, hence it was marginally superior.
- Future research may apply the suggested model to other crops using feature selection techniques. Lastly, adding more datasets could have given better predictions.

The yield of more crops will also be predicted, and crop data and soil characteristics will be combined to make a website or mobile app that simplifies crop selection. Although this is a very small step, we hope it is the start of something big. Hence, this model will be a powerful aid in helping the tea business make optimal management decisions early on in order to maximize profits from the estates.

REFERENCES

[1] Hossain, Md Akter, et al. "Predicting rice yield for Bangladesh by exploiting weather conditions." In: 2017 international conference on information and communication technology convergence (ICTC), pp. 589-594. IEEE, Jeju Iceland (2017).

[2] Chakrabarty, Amitabha, et al. "Prediction approaches for smart cultivation: a comparative study." *Complexity* 2021 (2021).

[3] Statista, https://www.statista.com, last accessed 2023/01/06.

[4] Wikipedia, https://en.wikipedia.org/wiki/Tea_production_in_Bangladesh, last accessed 2023/01/06.

[5] Business Inspection, https://businessinspection.com.bd/tea-industry-of-bangladesh, last accessed 2023/01/06.

[6] Ahamed, AT M. Shakil, et al. "Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh." In: *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 1-6. IEEE, Takamatsu, Japan (2015).

[7] Mamun, Shahriar, et al. "JuteBangla: A Comparative Study on Jute Yield Prediction using Supervised Machine Learning Approach based on Bangladesh Perspective." In: *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-5. IEEE, Kharagpur, India (2022).

[8] Ahmed, Selena, et al. "Global climate change, ecological stress, and tea production." In: *Stress Physiology of Tea in the Face of Climate Change*, pp. 1-23. Springer, Singapore (2018).

[9] Batool, Dania, et al. "A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning." In: *Plants* 11.15, p: 1925. MDPI, Switzerland (2022).

[10] Palanivel, Kodimalar, et al. "An approach for prediction of crop yield using machine learning and big data techniques." In: *International Journal of Computer Engineering and Technology* 10.3, pp. 110-118. SSRN, India (2019).

[11] BURHAN, Hasan Arda. "Crop Yield Prediction by Integrating Meteorological and Pesticides Use Data with Machine Learning Methods: An Application for Major Crops in Turkey." In: *Ekonomi Politika ve Finans Araştırmaları Dergisi* 7, pp. 1-18. IERFM Özel Sayısı, Ankara, Turkey (2022).

[12] Ahmed, Nisar, et al. "Development of Crop Yield Estimation Model using Soil and Environmental Parameters." In: *arXiv preprint arXiv:2102.05755*. Arixiv, New Maxico (2021).

[13] Jambekar, Suvidha, et al. "Prediction of crop production in India using data mining techniques." In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1-5. IEEE, Pune, India (2018).

[14] Nzuva, M. S., et al. "Prediction of tea production in Kenya using clustering and association rule mining techniques." In: *Am J Compt Sci Inform Technol* 5.2. iMedPub, UK (2017).

[15] Baruah, Rupanjali D., et al. "Use of data mining technique for prediction of tea yield in the face of climate change of Assam, India." In: *2016 International Conference on Information Technology (ICIT)*, pp. 265-269. IEEE, Bhubaneswar, India (2016).

[16] Das, Animesh Chandra, et al. "Integrating an expert system, gis, and satellite remote sensing to evaluate land suitability for sustainable tea production in bangladesh." In: *Remote Sensing* 12.24, p: 4136. MDPI, Dehradun (2020).

[17] Duncan, John MA, et al. "Observing climate impacts on tea yield in Assam, India." In: *Applied Geography* 77, pp. 64-71. Elsevier, China (2016).

[18] Phan, Phamchimai, et al. "Using multi-temporal MODIS NDVI data to monitor tea status and forecast yield: a case study at Tanuyen, Laichau, Vietnam." In: *Remote Sensing* 12.11, p: 1814. MDPI, Dehradun (2020).

[19] Jui, S. Janifer Jabin, et al. "Spatiotemporal Hybrid Random Forest Model for Tea Yield Prediction Using Satellite-Derived Variables." In: *Remote Sensing* 14.3, p: 805. MDPI, Dehradun (2022).

[20] Islam, Md Aliul, et al. "Fitting ARIMA model and forecasting for the tea production, and internal consumption of tea (per year) and export of tea." In: *International Journal of Material and Mathematical Sciences* 2.1, pp. 8-15. UniversePG, Malaysia (2020).

[21] Gandhi, Niketa, et al. "PredictingRice crop yield using Bayesian networks." In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 795-799. IEEE, Jaipur, India (2016).

[22] Schwalbert, Raí A., et al. "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil." In: *Agricultural and Forest Meteorology* 284, p: 107886. Elsevier, Australia (2020).

[23] Hammer, Ralph G., et al. "Sugarcane yield prediction through data mining and crop simulation models." In: *Sugar Tech* 22.2, pp. 216-225. Springer (2020).

[24] Khaki, Saeed, et al. "Crop yield prediction using deep neural networks." In: *Frontiers in plant science* 10, p:621. Frontiers (2019).

[25] Yethiraj, N. G. "Applying data mining techniques in the field of agriculture and allied sciences." In: *International Journal of Business Intelligents ISSN*, pp. 2278-2400. IIR, Chennai, India(2012).