Assessing Predictive Models for Tea Yield: A Statistical and Machine Learning Approach in Assam's Biswanath Chariali District

Original research article

Abstract

Climatic factors significantly impact Assam tea production. The tropical climate of Assam, characterized by high precipitation and temperatures up to 36°C during the monsoon, creates ideal conditions for tea cultivation, contributing to the region's unique malty flavor**.** Here, in this study an attempt has been made to bring a comparison among statistical and machine learning models in prediction of tea production and evaluate an optimal model among them. A time span of last 23 years data were collected from Biswanath College of Agriculture under Assam Agriculture University situated at Biswanath Charialidistrict. The study has found that mean absolute prediction error of random forest regression model is 6.49 percent followed by decision tree (7.3 percent) and linear regression model (7.5 percent).From the evaluation metrics, random forest algorithm fits well in comparison to decision tree and linear regression. This study could be generalized to comparison among more predictive machine learning models.

Keywords

Assam tea; prediction; machine learning; climatic factors

1 Introduction

Climatic factors significantly impact Assam tea production. The tropical climate of Assam, characterized by high precipitation and temperatures up to 36°C during the monsoon, creates ideal conditions for tea cultivation, contributing to the region's unique malty flavor [1]. Climate change poses challenges such as prolonged droughts, extreme temperatures, and increased pest infestations, affecting tea production in Assam and other regions like Dooars in West Bengal [2] [3] [4]. Adaptive strategies like rainwater harvesting, afforestation, and using climate-resistant cultivars are being adopted by tea growers to mitigate these impacts and ensure sustainable production [5]. Additionally, studies in Dooars have shown that temperature variations during different seasons, excessive rainfall, and changes in solar radiation and soil temperature can either positively or negatively influence tea yield, emphasizing the need for proactive measures to safeguard tea plantations from the adverse effects of climate change[20-22]. Statistical and machine learning techniques have been compared in predicting Assam tea production. Studies have shown that machine learning algorithms, such as XGBoost regressor and random forest models, outperform statistical methods like multiple linear regression in tea yield prediction [6] [7]. Additionally, the use of crop simulation models like AquaCrop and machine learning algorithms has been found to provide more accurate predictions with lower errors compared to traditional statistical approaches [8]. Furthermore, the selection of suitable sites for tea cultivation has been enhanced through the application of random forest models, emphasizing the importance of climate and soil factors in tea farming [9]. Moreover, the development of hybrid models like AOA-SVM has significantly improved soil moisture content prediction in tea plantations, showcasing the potential of machine learning in enhancing agricultural practices [10].In order to anticipate tea output in the Biswanath Chariali district, the present research has attempted to compare traditional statistical approaches with machine learning algorithms and assess which model or technique is appropriate. The use of machine learning algorithms in prediction of tea production of Biswanath Chariali district may increase predictive ability to certain extent in contrast to classical statistical models. Additionally, the study modelled many climate parameters associated with tea production that were reported at the Biswanath Chariali weather station.In this study, the observation of different weather factors on tea production over the years is important for its production growth and growing ability under certain conditions.The study's observations included the variances accounted for by several climate conditions.

2. Materials and Methods

2.1. Study area

The district of Biswanath, Assam, covers an area of 1100 square kilometres and is located on the north bank of the Brahmaputra River. It became a separate district on August 15, 2015, when it was divided from Sonitpur district. The area is located between longitudes 26°14'00" and 27°0'00" North and between longitudes 92°52'30" and 93°50'0" East. Assam Agricultural University (AAU)'s second constituent college, Biswanath College of Agriculture, was founded with the need for the general and comprehensive development of agriculture and related fields in mind, both for the state of Assam as a whole and for the entire north bank valley in particular. Beginning on February 2, 1988, the college offered an academic programme. In the Assamese North Bank Plains Zone, near Biswanath Chariali, the college was founded.

Map of the area under study

Figure 1: Map of Biswanath Chariali district

2.2 Data

The data for the study adopted is secondary in nature. Tea production data cultivated in Biswanath College of Agriculture (BNCA)under Assam Agriculture University situated at Biswanath Chariali district was collected from 2000 to 2023. The climatic data comprise of different factors such as rainfall, bright sunshine hours, relative humidity and maximum and minimum temperature werecollectedfor the study period respectively. The climatic variables used in the following study are discussed below:

- **Rainfall**

Tea production is a crucial sector in many regions, such as Darjeeling, Assam, or Sri Lanka [11]. The climate and soil conditions in these areas are ideal for growing tea. However, the success of tea production is heavily dependent on weather factors, particularly rainfall. Insufficient rainfall can have a negative impact on tea production, leading to lower yields and potentially poorer quality of the tea leaves.

- **Bright sunshine hours**

One specific weather factor that has a significant impact on tea production is the number of bright sunshine hours [12]. Research has shown that tea plants require at least 6 hours of bright sunshine per day for optimal production. This is because bright sunshine provides the necessary energy for photosynthesis, which is crucial for the growth and development of tea plants.

- **Relative humidity**

Tea production is a delicate process that is highly influenced by various environmental factors. Relative humidity plays a crucial role in determining the quality and quantity of tea production. Tea plants require a specific range of relative humidity levels to thrive and produce optimal yields. High relative humidity levels can lead to increased fungal growth and disease susceptibility in tea plants, impacting the overall health and productivity of the crop

- **Temperature**

Tea production is greatly influenced by various factors, including weather conditions such as temperature [12]. Temperature plays a crucial role in tea production, as it directly affects the growth and development of tea plants. Tea plants thrive in specific temperature ranges, and any deviation from these optimal conditions can have a significant impact on tea production.

2 Material and Methods

For analysis purpose, statistical and machine learning techniques were used. Most of the cause and effect research study has been considering a linear relationship among dependent and independent variables. This study has also assumed a linear relationship of tea production with other weather variables. To fulfillment of the objective, multiple linear regression model as statistical model and decision tree and random forest algorithms for machine learning counterparts were considered respectively.The analysis of the study is performed using

SPSS(version 22) software and python programming language [13]. Sklearn module is used for supervised machine learning regression algorithms. Seaborn module is used for data visualization. The following statistical and machine learning regression models are used in the study are discussed below:

- **Multiple Linear Regression**

To better understand the factors that influence tea production, multiple linear regression analysis is often used. By applying multiple linear regression analysis, researchers can identify the variables that have a significant impact on tea production. These variables may include factors such as rainfall, temperature, soil fertility, labor availability, and investment in agricultural technology. By examining the relationship between tea production and these independent variables, researchers can determine the extent to which each variable contributes to tea production. The following model is given as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots\ldots\ldots + \beta_n x_n \ldots\ldots\ldots\ldots\ldots\ldots \quad (i)$$

Where,

y = the predicted value of the dependent variable

β_0 = the y-intercept

$\beta_1 x_1$ = the regression coefficient (β_1) of first independent variable (x_1).

$\beta_n x_n$ = the regression coefficient (β_n) of n^{th} independent variable (x_n).

- **Decision tree**

The supervised learning method includes decision trees as one of its tools. Using decision nodes at each stage of the algorithm, the flow-chart looks like a tree structure. Results nodes are what remain at the end of the algorithm. Both classification and regression problems are addressed by the decision tree technique. Both continuous and discrete variable values are predicted by the algorithm through training in regression.

- **Random Forest**

Regression and classification tasks are handled by a collection of methods called random forest. The bootstrap and aggregation method, commonly referred to as bagging, is employed by the random forest algorithm. As opposed to the outcomes produced by individual decision

trees, the algorithm employs numerous decision trees to get a resolution on a given issue. Outliers and noisy features can be effectively detected with it.

3 Evaluation of the fitted models

The data of machine learning algorithms are split into two parts- one for training of the models and the other for testing as well as evaluation of the models performance. The training part and testing part of the data consists of 75 percent and 25 percent respectively.

In order to evaluate the performance of the selected statistical and machine learning regression models, following measures are used for evaluation of the fitted models:

- **R-Squared:**

  In a regression model, R-Squared (also known as R^2 or the coefficient of determination) is a statistical metric that establishes how much of the variance in the dependent variable can be accounted for by the independent variable[18]. The formula is given as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad \text{..... (ii)}$$

Where,

SS_{res}= residual sums of square

SS_{tot}= total sums of square

- **Mean Squared Error**

The value of mean squared error also contributes to the regression model evaluation[19]. It is used to measure the closeness of dispersion of the fitted regression line from the given data points. A high value of mean squared error will be an indicator for high dispersion of the observations from the regression line and vice versa. It is computed as

$$Mean\ Squared\ Error = \frac{\sum(y_i - \hat{y}_i)^2}{n} \quad \text{..... (iii)}$$

- **Mean Absolute Percentage Error**

In statistics, a forecasting method's prediction accuracy is measured by the mean absolute percentage error (MAPE), often referred to as the mean absolute percentage deviation (MAPD).

$$Mean\ Absolute\ Percentage\ Error\ = 100\frac{1}{n}\sum_{t=0}^{n}\left|\frac{A_t - F_t}{A_t}\right| \ldots\ldots\ldots\ldots\ldots\ldots \text{(iv)}$$

Where,

A_t = *Actual value.*

F_t = *Forecastvalue.*

4 Results

The descriptive statistics of tea production and climatic variables are given as follows:

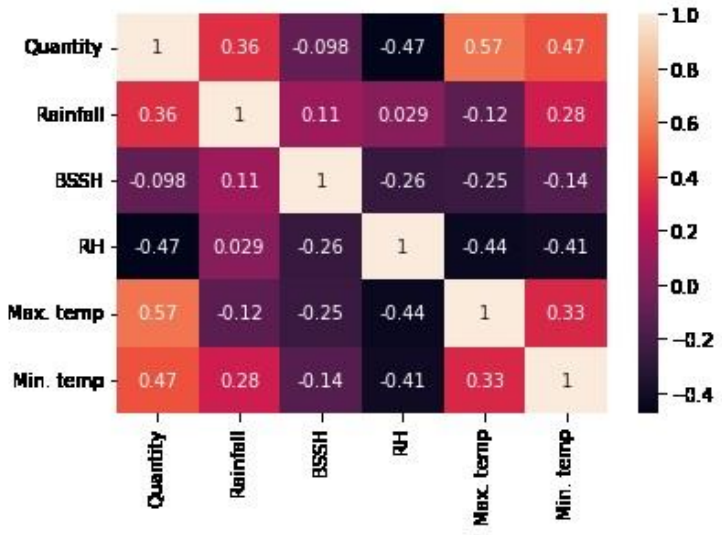
Table 1: Descriptive statistics of variables under study from 2000 to 2023

	Mean	Std. Deviation	Variance
Quantity	6819.46	2481.018	6155448.955
Rainfall	1879.8592	311.74929	97187.619
BSSH	64.8146	4.59636	21.127
RH(E)	62.9929	4.22475	17.848
Max. temp	347.8797	6.72813	45.268
Min. temp	219.5497	11.88799	141.324

Source: Statistical Package for Social Sciences

From table 1, the average tea production from 2000 to 2023 is 6819.46 and standard deviation 2481.018. The table 1 also depicts the mean and standard deviation of rainfall, BSSH, relative humidity in evening and temperature (maximum and minimum).

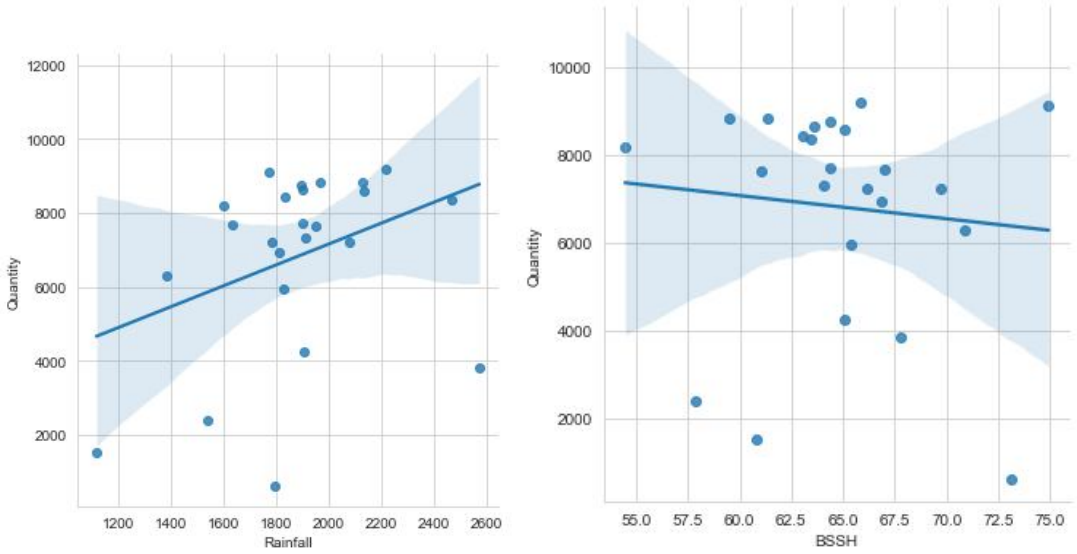
Figure 2: Graphical representation of Pearson correlation coefficient matrix



Source: Python programming language

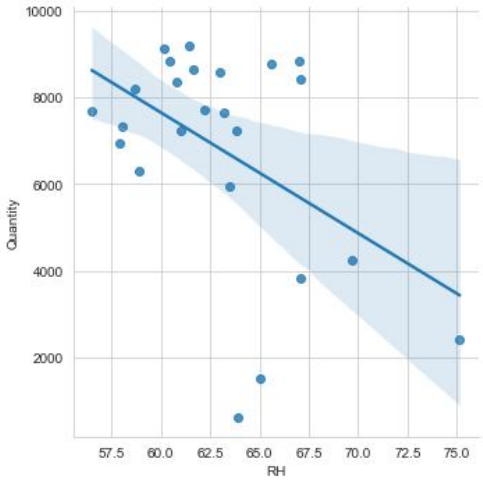
From figure 1, it could be observed that rainfall, maximum temperature and minimum temperature have a positive correlation with tea production whereas relative humidity and bright sunshine hours has a negative correlation.Since correlation is a measure of linear association, from the matrix it may be observed that if relative humidity and bright sunshine hours decreases, the tea production could increase or vice versa.

Figure 3: Graphical representation of regression plots of variables under study

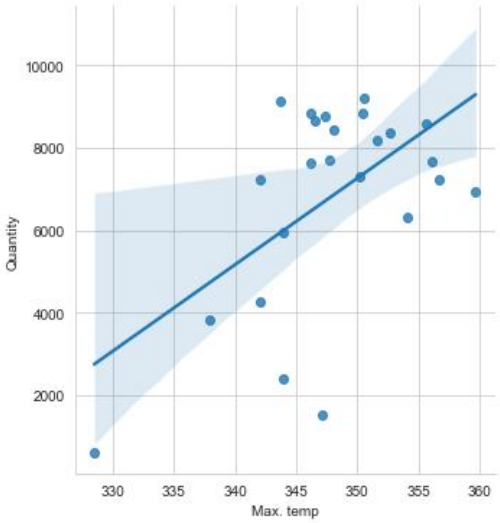


Source: Python programming language

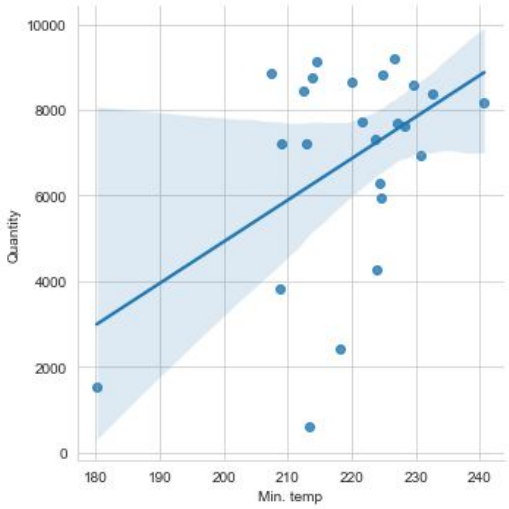
Source: Python programming language



Source: Python programming language



Source: Python programming language



Source: Python programming language

Table 2: Evaluation metrics of Statistical and Machine Learning Models

Models	Mean squared error	Mean absolute percentage error	R-squared
Linear Regression	3330530.26	0.0753	0.57
Decision Tree	424628.59	0.0732	058

Random Forest	362856.04	0.0649	0.64

Source: Python programming language and Statistical Package for Social Sciences

The r-square, mean absolute percentage error and mean squared error of the supervised machine learning and statistical regression models are displayed in Table 2. The random forest regression model has the lowest mean squared error value, followed by the multiple linear regression and decision tree models. Table 2 shows that the random forest regression algorithm's r-square value is higher than that of the other models.Further, from the table it could be observed that mean absolute percentage errors (MAPE) of the models are below 10 percent. The MAPE of random forest is 6.4 percent followed by decision tree and linear regression model.Since the mean squared value for a linear regression model is large, it suggests that the data points are spread apart from the fitted line. Compared to the linear regression model, the random forest algorithm has a lower value than the decision tree algorithm. This suggests that the observed values are closer to the regression line that the algorithm has fitted.

5 Discussion and Conclusion

Data mining tools were used to assess the influence of major climate variables such as temperature, rainfall, sunshine, and precipitation on tea yield in Assam, India, leading to the development of a crop yield forecast model based on weather variables using multiple linear regression [14]. Some studies were focused on analyzing factors influencing tea productivity in Northeast India through a combined statistical and modeling approach **[**15**]**. Research studies compared the performance of 1-D CNN with MLR and KNN models for chlorophyll estimation, demonstrating that leaf color features can effectively predict chlorophyll content [16]. Further the research studies contributes by exploring the application of classification techniques like random forest classifier, k-nearest neighbor classifier, support vector machine classifier, and neural network in the context of tea leaf disease prediction. These techniques can be valuable for practical implementation in the agriculture sector, specifically in tea cultivation[17]. The present research study was undertaken to bring a comparison among statistical and machine learning regression method in prediction of tea production of Biswanath Chariali district tea prodcution. From the analysis of the results, it could be concluded that random forest regression model fits well in comparison to decision tree and linear regression model. Further research could be carried out in consideration of more machine learning algorithms and neural network models could also be considered. Such

studies could bring more focus on studying the effect of weather factors on different agricultural production of the state with advanced algorithms.

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 writing or editing of manuscripts.

6. References

[1]. Babu, A., Roy, S., Baruah, R.D., Deka, B., Ahmed, K.Z., Bayen, S. and Sarkar, S. (2022). Impact of Climate Change on Tea Cultivation and Adaptation Strategies. In Climate Change and Agriculture, N. Benkeblia (Ed.). https://doi.org/10.1002/9781119789789.ch13

[2]. Pradip, Baruah., Gautam, Handique. (2021). Perception of climate change and adaptation strategies in tea plantations of Assam, India. Environmental Monitoring and Assessment, doi: 10.1007/S10661-021-08937-Y.

[3]. Piyashee, Mallik., Tuhin, Ghosh. (2021). Impact of climate on tea production: a study of the Dooars region in India. Theoretical and Applied Climatology, doi: 10.21203/RS.3.RS-276873/V1.

[4]. Pranab, Dutta., Himadri, Kaushik., R., P., Bhuyan., Pranjal, Kr, Kaman., Arti, Kumari., Apurba, Das., H., K., Saikia. (2020). Relation of Climatic Parameter on Tea Production in Organic Condition Specific to Assam. International Journal of Current Microbiology and Applied Sciences, doi: 10.20546/IJCMAS.2020.904.269.

[5]. Piyashee, Mallik., Tuhin, Ghosh. (2021). Impact of surface-net solar radiation and soil temperature on tea production in India: a study of the Dooars region in West Bengal. Regional Environmental Change, doi: 10.1007/S10113-021-01844-5.

[6]. Dania, Batool., Muhammad, Shahbaz., H., Shahzad, Asif., Kamran, Shaukat., Talha, Mahboob, Alam., Ibrahim, A., Hameed., Zeeshan, Ramzan., Abdul, Waheed., Hanan, A., Aljuaid., Suhuai, Luo. (2022). A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning. Plants, doi: 10.3390/plants11151925.

[7]. Netrananda, Sahu., Pritiranjan, Das., Atul, Saini., Suraj, Kumar, Mallick., Rajiv, Nayan., S., P., Aggarwal., Balaram, Pani. (2023). Analysis of Tea Plantation Suitability Using Geostatistical and Machine Learning Techniques: A Case of Darjeeling Himalaya, India. Sustainability, doi: 10.3390/su151310101

[8]. Dongxiao, Yin., Yanhua, Wang., Ying, Huang. (2023). Predicting soil moisture content of tea plantation using support vector machine optimized by arithmetic optimization algorithm. Journal of Algorithms & Computational Technology, doi: 10.1177/17483026221151198,

[9]. Isabel, R., Fulcher. (2022). Prediction of Crops Production Using Random Forest Regression. doi: 10.1007/978-981-19-1657-1_8.

[10]. Ying, Huang. (2023). Improved SVM-Based Soil-Moisture-Content Prediction Model for Tea Plantation. Plants, doi: 10.3390/plants12122309.

[11]. Dutta, P., Kaushik, H D., Bhuyan, R., Kaman, P K., Kumari, A., Das, A., & Saikia, H K. (2020). Relation of Climatic Parameter on Tea Production in Organic Condition Specific to Assam.

[12]. Redden, R J., Hatfield, J L., Prasad, P V., Ebert, A W., Yadav, S., & O'Leary, G. (2013). Temperature, climate change, and global food security. , 181-202. https://doi.org/10.1002/9781118308240.ch8.

[13]. Garreta, R., &Moncecchi, G. (2013). Learning scikit-learn: Machine Learning in Python. http://cds.cern.ch/record/1641744.

[14]. Rupanjali, D., Baruah., Sudipta, Singha, Roy., R., M., Bhagat., L., N., Sethi. (2016). Use of Data Mining Technique for Prediction of Tea Yield in the Face of Climate Change of Assam, India. 265-269. doi: 10.1109/ICIT.2016.060.

[15]. Rishiraj, Dutta., Eric, Smaling., Rajiv, Mohan, Bhagat., Valentyne, A., Tolpekin., Alfred, Stein. (2012). Analysis of factors that determine tea productivity in northeastern india: a combined statistical and modelling approach. Experimental Agriculture, 48(1), 64-84. doi: 10.1017/S0014479711000834.

[16]. Utpal, Barman. (2021). Deep Convolutional neural network (CNN) in tea leaf chlorophyll estimation: A new direction of modern tea farming in Assam, India. Journal of Applied and Natural Science, 13(3), 1059-1064. doi: 10.31018/JANS.V13I3.2892

[17]. Alok, Ranjan, Srivastava., M., Venkatesan. (2020). Tea Leaf Disease Prediction Using Texture-Based Image Processing. 17-25. doi: 10.1007/978-981-15-0135-7_3

[18]. Glantz, Stanton A.; Slinker, B. K. (1990). Primer of Applied Regression and Analysis of Variance. McGraw-Hill. ISBN 978-0-07-023407-9.

[19]. Hyndman, Rob J., and Anne B. Koehler (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

[20]. Harshith, Maddila, Ayushi Sahu, Sanju Indrakanti, R. Kameshwar Reddy, and Sunil Bhutada. 2023. "Optimizing Crop Yields through Machine Learning-Based Prediction".

Journal of Scientific Research and Reports 29 (4):27-33. https://doi.org/10.9734/jsrr/2023/v29i41741.

[21]. Pangarkar, Darshan Jagannath, Rajesh Sharma, Amita Sharma, and Madhu Sharma. 2020. "Assessment of the Different Machine Learning Models for Prediction of Cluster Bean (Cyamopsis Tetragonoloba L. Taub.) Yield". Advances in Research 21 (9):98-105. https://doi.org/10.9734/air/2020/v21i930238.

[22]. Van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture. 2020 Oct 1;177:105709.