Machine Learning Techniques for Sugarcane Yield Prediction Using Weather Variables

Ali J. Ramadhan[1*], *S. R.* Krishna Priya[2], *V* Pavithra[2], *Pradeep* Mishra[3], *Abhiram* Dash[4], *Mostafa* Abotaleb[5], *Hussein* Alkattan[5] and *Zainalabideen* Albadran[1]

[1]University of Alkafeel, Najaf, Iraq
[2]PSG College of Arts and Science, Tamil Nadu, India
[3]JNKVV College of Agriculture, Rewa, India
[4]Orissa University of Agriculture and Technology, Odisha, India
[5]South Ural State University, Chelyabinsk, Russia

Abstract. Weather has a profound influence on crop growth, development and yield. The present study deals with the use of weather parameters for sugarcane yield forecasting. Machine learning techniques like K- Nearest Neighbors (KNN) and Random Forest model have been used for sugarcane yield forecasting. Weather parameters namely maximum temperature and minimum temperature, rainfall, relative humidity in the morning and evening, sunshine hours, evaporation along with sugarcane yield have been used as inputs variables. The performance metrics like R^2, Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) have been used to select the best model for predicting the yield of the crop. Among the models, Random Forest algorithm is selected as the best fit based on the high R^2 and minimum error values. The results indicate that among the weather variables, rainfall and relative humidity in the evening have significant influence on sugarcane yield.

1 Introduction

India is the largest producer of agricultural goods. In our country, farming has been around for a long time. Farming can be subsistence or commercial, depending on the sort of farm. Food crops are grown in different regions of the country due to soil, climate, and cultivating practices that vary from region to region. Rice, wheat, sugarcane, and other staple crops are common in India. One of India's most important cash crops, sugarcane has a significant impact on both the agricultural and industrial economies [1]. It is grown on 5.0 million hectares of land and produces 350 million tons annually. India's tropical and subtropical regions both have sugarcane plantations. Sugarcane yield is determined by the weather throughout the crop growing period [2]. During the growing season, sugarcane productivity

* Corresponding author: ali.j.r@alkafeel.edu.iq

is highly dependent on a unique succession of everyday meteorological conditions. It is a long-lived crop, thus it experiences all three seasons during its life cycle: rainy, winter, and summer. Temperature, light, and moisture availability are the primary climatic factors that influence cane growth, production, and quality [4]. Crop yields are heavily influenced by the environment and weather conditions they are exposed to. Crops may be benefited or harmed by extremely low temperatures. Due to the fact that crops can be damaged by both low and high rainfall in India, a consistent amount of rain is needed to ensure good agricultural output. Other elements, such as temperature and humidity, also have an impact on crop productivity. Climate is a major determinant of crop yield that is out of our control [5]. Using meteorological variables to model and forecast agricultural yields has thus become an intriguing study area. Policymakers and researchers can use crop yield forecasting models based on meteorological parameters to build sustainable cropping plans [6].

Application of machine learning techniques in agriculture would help in accurate prediction crops of yield.

2 Background

This section summarizes how machine learning techniques can be applied for crop production forecasting, as well as related publications in the literature on this rapidly increasing research field.

2.1 Machine learning for crop yield forecasting

Decisions about what crops to cultivate and how to grow them can be supported by machine learning (ML), which is a critical decision support tool for agricultural yield prediction.

When the relationship between the input and output variables is uncertain or difficult to determine, machine learning has to cope with it. Structural descriptions can be learned automatically from examples using the term "learning". Data model assumptions aren't made in ML, unlike in traditional statistical methods. Non-linear phenomena like agricultural yield forecasts can benefit from this feature. A common application of ML is in predicting crop yields. In machine learning, there are four main forms of learning: supervision, un supervision, semi-supervised, and reinforcement learning. A target variable (or dependent variable) is predicted by the Supervised Learning algorithm based on a collection of predictors (independent variables). Using these variables, we can build a function that converts inputs to outputs and back again. To achieve the necessary level of accuracy on training data, the model is trained repeatedly. Regression, decision trees, random forests, KNN, and logistic regression are all examples of supervised learning [7].

2.2 Related works

Machine learning for crop yield prediction is a new trend. Plot yield prediction using machine learning methods.

To forecast agricultural yield using climate information, machine learning was applied. Rainfall, maximum and minimum temperatures, potential evapotranspiration, cloud cover, and rainy day were used [8]. Rice yield forecast considering annual rainfall, food price index, and irrigated area [9]. The models of quadratic, linear, polynomial, and stepwise linear regression were compared [10]. These algorithms were tested on cotton, sugarcane, and turmeric yields [11]. Data mining was utilized to predict crop yields [12]. The gradient booster tree algorithm was shown to be more accurate [13]. In this study, meteorological variables were used to predict agricultural productivity. The study's findings suggested better prediction accuracy [14]. The performance of SVM, decision tree, and lasso regression for sugarcane yield prediction was examined [15], employed an association rule mining system to estimate agricultural yield. The algorithm contains Apriori, Elcat, and AprioriTid. These algorithms predicted the most profitable crops and their probable price at harvest [16]. Sugarcane and rice yield prediction using regression technique [17, 18]. Algorithms for

cotton and sugarcane crop forecasting were compared by [19, 20]. Among the three machine learning algorithms tested, random forest outperformed polynomial regression and decision trees [21].

In this work, KNN and Random Forest models were tested for sugarcane yield forecasting in Coimbatore.

3 Description of data

The study has been conducted in Coimbatore district of the State of Tamil nadu.

3.1 Data on weather parameters

Daily data on weather parameters such as maximum temperature (^{0}C), minimum temperature (^{0}C), relative humidity 7 hrs (%), relative humidity 14 hrs (%),rainfall (mm), evaporation (mm/day), sunshine (hours). Daily Weather data has been collected from AgroMetrology Department of Tamil Nadu Agricultural University for a period of 56 years (1960-2016).

3.2 Sugarcane yield data

The Sugarcane data (tonnes/ hectare) for Coimbatore District has been collected from different volumes of Annual, Season and crop report issued by Directorate of Economics and Statistics issued by State Government of Tamil Nadu

4 Methodology

4.1 Data understanding

The simplest and most important component, to any data experiment is data understanding.Before applying any machine learning algorithm, complicated tools and advanced statistics,proper understanding of data is essential.Data understanding gives an idea about the first look at the data. Analyzing data to discover patterns, anomalies, and test hypotheses using summary statistics and graphs is known as data mining. To get the most out of your data, it's best to examine it first. The problem was described and the data set was defined using summary statistics from the current data set.

4.2 Data pre-processing

Preparing data for use in the training of a machine learning model is known as data preprocessing. Data preparation is another name for this process. Until the data is ready to be used, it cannot be analyzed effectively. Possible reasons for this include faulty preparation or omissions in the data. The phrase "pre-processing" refers to the steps taken before data is delivered to the algorithm to make changes. It's a method for transforming raw data into a clean dataset.

4.3 Need for data pre-processing

Data preprocessing helps the machine learning model produce better results, increasing the efficiency and accuracy of the model.

Data should be prepared so that many machine learning algorithms can run on the same data set and the best one is selected as another consideration.

The sklearn.preprocessing, a pre-built capability in the Scikit-learn module, was used in the current inquiry because it is pre-built in Python.

4.3.1 Steps in data preprocessing

- Importing the libraries
- Importing the dataset
- Taking care of missing data and outliers
- Checking for Multi-Collinearity.
- Splitting the data into test and train

4.3.2 Checking for multi-collinearity

The tight connection between two explanatory variables is called collinearity. Multi-collinearity occurs when two or more explanatory factors are correlated. Multi-collinearity affects linear regression model interpretability. On can use the linear regression model to establish not only whether a response is linked to an explanatory component, but also their individual impacts. Multi-collinearity obscures specific impacts.

4.3.3 Heat map of correlations

For multi-collinearity, heat maps can be used to identify its presence. By modifying the color of positive and negative correlation and the size of magnitude, a heatmap of correlations helps to better visualize the data.Heat map is a two-dimensional array and is used to visualize variations across categories. It is colorful pictured graph, the darker the color it shows negative correlation and lighter the color shows the positive correlation. Heat map of correlation is presented in figure.

4.3.4 Splitting the data into test and train

In order to evaluate the performance of machine learning algorithms, we can divide the data into test and training data sets, and then utilize the test data to make predictions.

Machine learning toolkit, called Scikit-learn, is used to implement training and testing in two phases. An 80:20 split was employed in this study, which implies that 80% of the data was used for training, while 20% was used to test the model created from it.

Data from 1960 to 2011 have been used for training set and 2012-2016 have been used for testing set. Here, Sugarcane yield is a x variable which is dependent variable. And trend value, Maximum temperature, minimum temperature, sunshine, evaporation, rainfall, related humidity 7, related humidity 14 are y variable which is independent variables or explanatory variables. The steps involved in machine learning are shown in Figure 1 below.

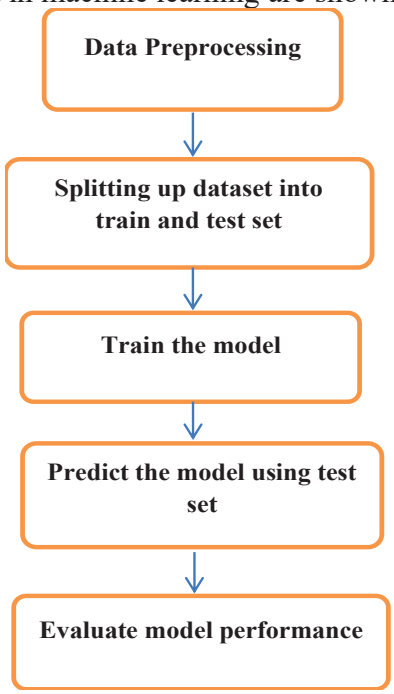

Fig. 1. The steps involved in machine learning

4.4 K-Nearest neighbor (KNN) algorithm

There is a simple machine learning algorithm called K-Nearest Neighbor that uses the Supervised Learning technique. In the KNN approach, it is assumed that the new case/data is comparable to existing cases, and the new case is placed in the category that is most similar to the existing ones. If you need to add missing values or resample data, you can use this

adaptable approach. K-Nearest Neighbors (data points) are used to predict the class or continuous value of a new data point, much like the name implies. In other words, KNN is not based on any assumptions about the data it uses. Regression and classification can both benefit from KNN. The regression algorithm used in this study was KNN. Due to its inability to learn from the training data right away, it is sometimes referred to as a "lazy learner algorithm." Instead, it keeps the data set and later uses it to perform classification operations. Instead of learning weights from training data to predict output as in model-based algorithms, the KNN algorithm uses full training instances to predict output for unknown data. This is known as "instance-based learning."

4.5 Random forest algorithm

Random decision forest, a supervised learning technique, is one of several options. A "forest" of decision trees is created by "bagging," a frequent method of teaching. It's a technique for integrating several kinds of learning models in order to get better outcomes. Simply put, random forest is one of the most useful methods for machine learning. It's a popular algorithm because it's simple and versatile. Random Forest uses a combination of several different decision trees to get a more accurate forecast. Random forests, for example, are commonly used in modern machine learning systems to address classification and regression problems. In our investigation, we used the random forests technique.

During the process of building the trees, Random Forest introduces more randomness. Instead than using the most significant characteristic to divide a node, a random subset of features is utilized to choose the best one. This leads in a more diverse model.

Relative importance using random forest algorithm

Additionally, the random forest technique makes it simple to examine the relative importance of each attribute in the forecast.. Python's scikit-learn machine learning library is a powerful tool. In Scikit-learn, there is a nice method for evaluating the usefulness of a feature by looking at how much impurity it reduces across the forest. Once a feature has been trained, this score is calculated for that feature and scaled to one. We used the Sklearn toolkit to determine the relative relevance of meteorological variables in sugarcane yield.

4.6 Performance metrics

The implemented KNN and Random Forest algorithm have been evaluated by using performance metrics which includesR – Squared value, Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE).

R- Squared (R^2) $$= 1 - \sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2 \quad (1)$$

Mean Absolute Error (MAE) $$= \sum |y_i - \hat{y}_i| / n \quad (2)$$

Mean Squared Error (MSE) $$= \sum (y_i - \hat{y}_i)^2 / (n-p) \quad (3)$$

Root Mean Square Error (RMSE) $$= [\sum (y_i - \hat{y}_i)^2 / n]^{1/2} \quad (4)$$

Mean Absolute Percentage Error $$= 1/n \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \quad (5)$$

Where '*n*' denotes total number of observed values and '*p*' denotes the number of model parameters.

5 Results and discussions

Correlation matrix is an important tool of exploratory data analysis. Correlation is to check whether the two variables correlate or not. Correlation can be done using heat maps. The Figure 2 shows the pictorial representation of correlation matrix using heat map.

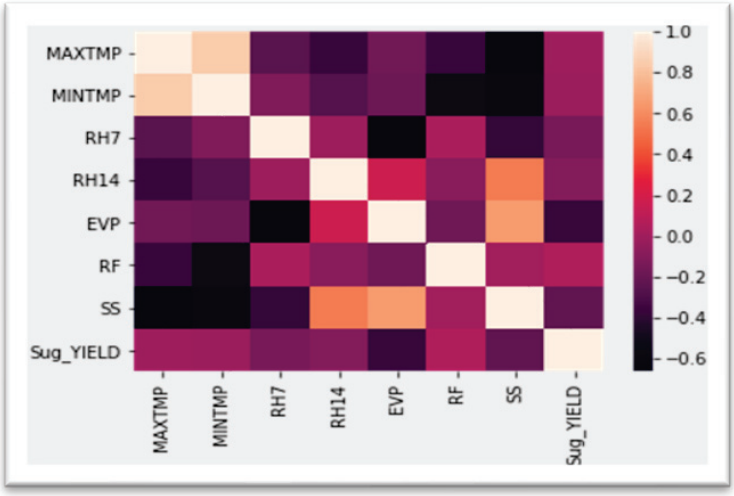

Fig. 2. Correlation matrix

Table 1. Performance metrics of KNN for K = 3,4,5,6,7

Neighbors	R^2	MSE	MAE	RMSE	MAPE
3	0.62	188.75	11.12	13.74	11.59
4	0.56	180.95	9.97	13.45	10.57
5	0.53	178.04	9.88	13.34	10.48
6	0.44	197.05	10.42	14.03	11.054
7	0.44	279.82	13.31	16.73	13.93

The above table 1 shows the values of performance metrics of KNN model. In the table it shows that K=5 is a best fit when compared to other neighbors. The best fit is concluded based on MSE, MAE, RMSE and MAPE values. For K=5 the error values are minimum.

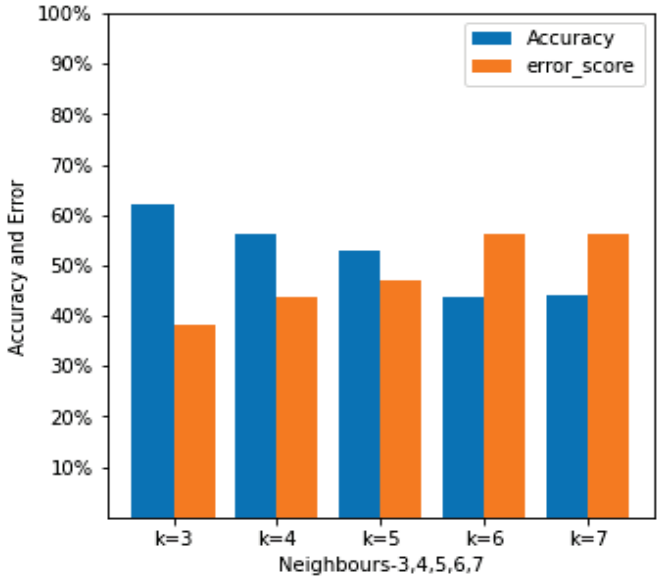

Fig. 3. Comparison of Accuracy and Error score of K= 3,4,5,6,7

The above figure 3 shows the accuracy and error rate of KNN model. The above model clearly shows that accuracy decreases and error increases as 'k' value increases. The below figure 4 shows the accuracy graph of KNN.

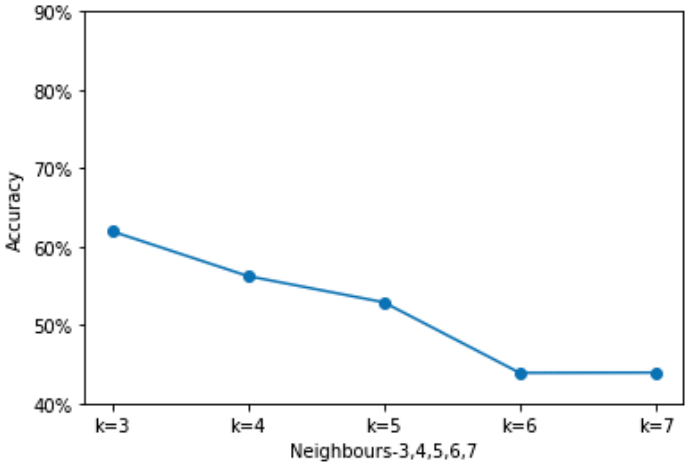

Fig. 4a. Accuracy score of K- Nearest Neighbors = 3,4,5,6,7

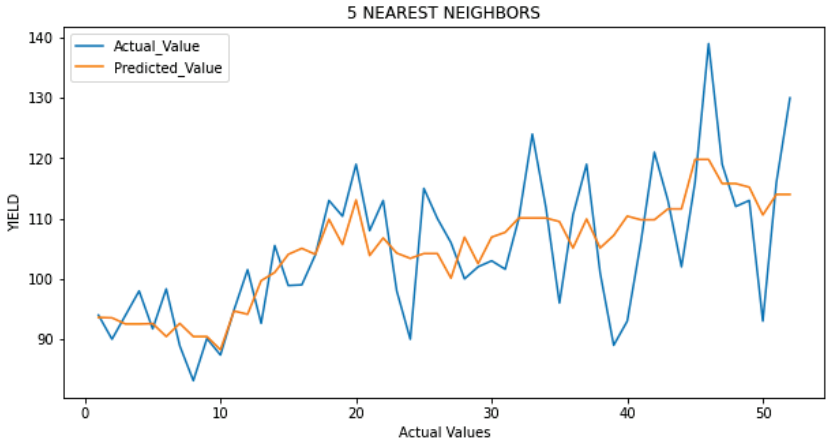

Fig. 4b. Comparison of Actual and Predicted values of KNN

Table 2. Performance metrics for Random Forest Regressor Algorithm

R2	0.93
MSE	145.99
MAE	10.53
RMSE	12.09
MAPE	10.72

The above table 2 shows that R square value is 0.923 which is best fit and Mean Absolute Percentage Error (MAPE) is 10.720 which is low which indicates that the chances of error is 10%.

Table 3. Relative importance of the predictor variable

Variable	Relative Importance
RF	0.13
RH14	0.11
SS	0.09
EVP	0.06
MINTMP	0.07
MAXTMP	0.06
RH7	0.04

The above table 3 shows the relative importance of variables, and it clearly shows that RF - Rainfall has the high relative importance followed by RH14- Relative humidity 14hrs in the evening and sunshine hours. And least relative importance is RH7 – Related humidity 7hrs in the morning.

6 Conclusion

The present investigation has been undertaken to identify the best model and to predict the yield of the sugarcane crop of Coimbatore district. The yield of sugarcane in the Random Forest model is identified as best fit having higher R^2value and least error values in comparison with KNN model. In the present study for the Random Forest model, MAPE value are found to be minimum with is 10% error. Using the Random Forest model, the relatively important weather variables in contributing to sugarcane yield have also been identified. Among the weather variable, rainfall, relative humidity at 14hrs and sunshine hours are important variables contributing to sugarcane yield. The random forest model can be successfully to forecast sugarcane yield of Coimbatore district.

References

1. Adil, N., Dewangan, S., Sharma, K.,2019,“Efficient Classification and Regression Techniques to Predict Crop Yield”, International Journal of Scientific and Technology Research,**8**,11,378-382.
2. Bhatla, R., Dani, B., Tripathi, A., 2018,“Impact of Climate on Sugarcane Yield over Gorakhpur District U.P using Statistical Model”,Vayu Mandal, **44**,1.
3. Ehsan khodadadi, S. K. Towfek, Hussein Alkattan. (2023). Brain Tumor Classification Using Convolutional Neural Network and Feature Extraction. Fusion: Practice and Applications, **13**(2), 34-41.
4. Everingham, Y., Sexton, J., Skocaj, D., Bamber G,I.,2016,“Accurate prediction of Sugarcane Yield Using a Random Forest Algorithm”,https://hal.archives-ouvertes.fr/hal-01532457.
5. Fathima, K., Sowmya, Basker, S., Kulkarni, S.,2020,“Analysis of Crop Yield Prediction Using Data Mining Technique”. International Research Journal of Engineering and Technology, **7**, 57708-7713.
6. Jakaria, A.H.M., Hossain, M.M, Rahman, M.A., 2020,“Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee”.arXiv preprint arXiv:2008.10789.
7. Josephine, B.M., Ramya, K.R., Rao, K.V.S.N.R., Kuchibhotla, S., Kishore,P.B.V., Rahamathulla, S.2020. “Crop Yield Prediction Using Machine Learning”, International Journal of Scientific & Technology Research, **9**, 2, 2102-2106.
8. Kiran, D.B., Priyanka, J., Poojitha, K.S., Khan, A.,2020,“Crop Yield Prediction Using Regression”, International Research Journal of Engineering and Technology,**7**, 5, 3896-3899.
9. Akbari, E., Mollajafari, M., Al-Khafaji, H. M. R., Alkattan, H., Abotaleb, M., Eslami, M., & Palani, S. (2022). Improved salp swarm optimization algorithm for damping controller design for multimachine power system. IEEE Access, **10**, 82910-82922.
10. Kumar, S., Singh, J., Singh, P.K., Pandey, D.K., 2018, “CoLK09204 (Ikshu-3) a new midlate maturing high yielding Sugarcane variety for Northwest Zone of subtropical India”,Indian Journal of Sugarcane Technology, **33**, 01, 39-43.
11. Kumar, M. N., 2018, Sugarcane Crop Yield Estimation Using K- Nearest Neighbors, Journal of Advanced Research in Dynamical & Control Systems, **10**, 04, 199-207.
12. Priya, P., Muthaiah, U., and Balamurugan, M., 2018,“Predicting yield of the Crop Using Machine Learning Algorithm”, International Journal of Engineering Sciences and Research Technology, **07**, 04, 1-7.

13. Sangeeta., Shruthi, G., 2020,"Design and Implementation of Crop Yield Prediction Model in Agriculture", International Journal of Scientific and Technology Research, 08, **01**,544-549.
14. Al-Mahdawi, H. K., Albadran, Z., Alkattan, H., Abotaleb, M., Alakkari, K., & Ramadhan, A. J. (2023, December). Using the inverse Cauchy problem of the Laplace equation for wave propagation to implement a numerical regularization homotopy method. AIP Conference Proceedings (Vol. 2977, No. 1). AIP Publishing.
15. Shastry, A., Sanjay, H.A., and Bhanushree, E.,2017," Prediction of Crop Yield Using Regression Techniques",International Journal of Soft Computing,**12**, 02, 96-102.
16. Sridhara, S., Ramesh, N., Gopakkali, P., Das, B., Venkatappa, S.D., Sanjivaiah, S.H., Singh, K.M., Singh, P., Ansary, D.O.E., Mahmoud, E.A., Elansary, H.O., 2020,"Weather -Based Neural Network, Stepwise Linear and Sparse Regression Approach for Rabi Sorghum Yield Forecasting of Karnataka, India", Agronomy, **10**, 1645.
17. Surya, P., Aroquiaraj, I.L., 2018,"Crop Yield Prediction in agriculture Using Data Mining Predictive Analytic Techniques",International Journal of Research and Analytical Reviews, **05**, 04, 783-787.
18. Al-Nuaimi, B. T., Al-Mahdawi, H. K., Albadran, Z., Alkattan, H., Abotaleb, M., & El-kenawy, E. S. M. (2023). Solving of the inverse boundary value problem for the heat conduction equation in two intervals of time. Algorithms, **16**(1), 33.
19. Veenadhari, S., Misra, B., Singh, C.D.,2014, "Machine Learning Approach for Forecasting Crop Yield Based on Climatic Parameters", International Conference on Computer Communication and Informatics.**03**, 05, 1-6.
20. Vishwa,G., Venkatesh, J., Geetha, C., 2019,"Crop Variety Selection Method Using Machine Learning".International Journal of Innovations in Engineering and Technology.**12**, 04, 35-38.
21. Wickramasinghe, L., Weliwatta, R., Ekanayake, P., Jayasinghe, J., 2021,"Modeling the Relationship between Rice Yield and Climate Variables Using Statistical and Machine Learning Techniques", Journal of Mathematics.Vol.**2021**, Article ID 6646126, 9 pages, 2021. https://doi.org/10.1155/2021/6646126