# ANALYSIS OF DIFFERENT REGRESSION TECHNIQUES FOR CROP YIELD PREDICTION

**Jaspreet Kaur[1*], Er. Lal Chand[2]**

**Abstract:** Machine learning is a crucial decision-support tool for predicting crop yields, enabling choices about which crops to cultivate and what exactly to do while they are in the developing season. The research on agricultural production prediction has been supported by the application of several machine-learning techniques. Machine learning algorithms can assist farmers in choosing which crop to cultivate in addition to boosting output by considering a variety of factors. The use of yield estimation by farmers can help them minimize crop loss and get the best prices for their produce. In this paper, agricultural yield per acre is predicted using. The paper compares multivariate polynomial regression, support vector machine regression, and random forest using RMSE, MAE, median absolute error, and R-square values.

**Keywords:** Dataset, Agriculture yield, Linear regression, XGB and Polynomial regression model.

[1*, 2]Department of Computer Sciences, Punjabi University, Patiala, Punjab

***Corresponding Author:** Jaspreet Kaur
*Department of Computer Sciences, Punjabi University, Patiala, Punjab

**DOI:** 10.53555/ecb/2022.11.12.340

## 1. Introduction

Agriculture has long been seen as India's primary and dominant cultural practice. Since ancient people farmed their own land, their requirements have been met. As a result, natural crops are grown and used by numerous organisms, including humans, animals, and birds. The greens products made on the land that the critter ate lead to a healthy and happy life. Since the development of new cutting-edge technology and methods, agriculture has been slowly declining. Due to these numerous inventions, individuals have focused on creating artificial and hybrid products, which can result in an unhealthy lifestyle [1]. There are no suitable solutions or technologies to deal with the predicament we confront after analyzing all these challenges and problems, including weather, temperature, and numerous other elements. There are numerous approaches to boost economic growth in the agricultural sector in India.

Crop quality and quantity can both be in creased and improved in a variety of way s. Additionally helpful for forecasting crop yield development is data mining [2].

Data mining is, in general, the process of analyzing data from several angles and condensing it into helpful information. Using data mining tools, users can categorize and summarize the associations discovered after analyzing data from a variety of dimensions or viewpoints. In big relational databases, data mining technically refers to the act of identifying correlations or patterns among numerous fields [3]. All these data's patterns, correlations, and interactions may include information. Information can be transformed into knowledge about past trends and potential developments. Accurate predictions enable better planning and decision-making regarding crop selection, resource allocation, risk management, and market strategies. This empowers stakeholders to optimize their practices, maximize productivity, and minimize losses. They can handle the complexity of interactions, leverage data-driven approaches, adapt to changing conditions, scale efficiently, improve performance over time, and provide decision support. These models enhance the understanding of the factors influencing crop yield and help stakeholders make informed decisions for better agricultural outcomes.

## 1.2 Related works

Numerous studies are being conducted to increase agricultural planning accuracy. The presented research objective is to increase accuracy as much as feasible. Numerous classification techniques are also employed to provide a fair percentage of crop production. The prior similar studies that other researchers have attempted are discussed in this section.

According to Zhang et al. [4], researchers consider the linear regression model while keeping in mind that crop yield predictions are frequently made using the Ordinary Least Square (OLS) estimation method. With a greater R, the autoregressive model outperformed the OLS in this case. The study found that ignoring temperature, NDVI and moisture correlated more to the wheat production in Iowa.

Qaddown and Hines [5] expand on the traditional regression neural networks. In order to anticipate tomato production in a growing environment, the Vapour Pressure Deficit (VPD), CO2, electromagnetic radiation, and temperature are taken into account in the Conformal Prediction (CF) architecture. This technique required the utilization of over 60,000 records.

Researchers Sanchez et al. [6] demonstrate a correlation between linear and nonlinear techniques for agricultural yield prediction. The comparison is performed using the best property subset for each method identified using split percentage validation and an entire algorithm from the preparation dataset. The technique takes the oldest information to build the models and then searches the training datasets for the best attribute subset. Unseen samples make up the test information sets, where performance is assessed. The most wellknown information-driven techniques for agricultural yield prediction, including stepwise linear regression, multiple linear regression, regression trees, and neural networks, were evaluated.

Rakesh Kumar et al. [7] concluded that this work aids in enhancing agricultural production rates by using various categorization techniques and contrasting various factors. To estimate agricultural yield, different machine learning techniques were evaluated. Artificial neural networks, support vector machines, decision trees, random forests, gradient-boosted decision trees, regularized greedy forests, and the proposed CSM technique (Crop Selection Method). The forecast is based on numerous parameters; thus, the accuracy and performance of the system vary depending on the parameters, the author concluded.

For repeated measures data, Ngaruye et al. [8] used Small Area Estimation (SAE) approaches to create district-level estimates of crop production for beans (i.e., bush and climbing beans) in Rwanda during the 2014 growing seasons. The Seasonal Agricultural Survey (SAS) 2014 microdata of the NISR were used by the authors to conduct their research.

According to Safieh et al. findings [9], climate change will likely have an impact on crop water needs as well as crop yields in the future. According to a study on the effects of harsh weather on various parts of Europe, rainfall and air temperature thresholds are the most accurate weather indicators of agricultural production. The two climatic parameters that are most frequently employed in studies are precipitation and air temperature. To predict crop yields, several factors have been employed, including solar radiation, air humidity, soil moisture, and wind speed. They used various machine learning models such as KNN, ANN, RNN, CNN etc. to compare its efficiency.

### 1.3 Contribution of the research

Machine learning models are useful in predicting crop yield for several reasons [10-11]:

**a)** Increased accuracy: Machine learning models can analyze large amounts of data and learn complex patterns that may not be apparent through traditional statistical methods. By considering various input features such as weather conditions, soil characteristics, historical yield data, and crop management practices, machine models can make accurate predictions of crop yield.

**b)** Improved Decision Making: Accurate crop yield predictions enable farmers and policymakers to make informed decisions regarding crop management, resource allocation, and market planning. By knowing the expected yield in advance, farmers can optimize irrigation, fertilizer application, pest control, and harvesting schedules. This helps to maximize productivity, reduce waste, and make better use of resources.

**c)** Risk Management: Crop yield prediction models can assist in risk management by providing insights into potential crop failures or yield fluctuations. Farmers can use this information to implement risk mitigation strategies such as diversification of crops, crop insurance, or adjusting planting schedules. By being prepared for potential yield variations, farmers can minimize losses and maintain stable income.

**d)** Sustainability and resource efficiency: By accurately predicting crop yield, farmers can optimize the use of resources such as water, fertilizers, and pesticides. They can avoid overuse or underuse of resources, leading to more sustainable agricultural practices. This can help reduce environmental impact, conserve resources, and improve overall efficiency.

**e)** Early warning system: Machine learning models can be used to develop early warning systems for crop diseases, pests, or extreme weather events. By monitoring various data sources and detecting patterns, these models can alert farmers in advance, allowing them to take timely preventive measures and mitigate potential damage to crops.

**f)** Planning and market analysis: Accurate crop yield predictions can help in planning agricultural activities, such as production forecasts and market analysis. This information is valuable for supply chain management, price determination, and market positioning. It can assist farmers, traders, and policymakers in making informed decisions related to crop production, storage, distribution, and market strategies.

Overall, machine learning models provide a data-driven approach to crop yield prediction, enabling farmers to optimize their practices, minimize risks, and make more informed decisions. This can lead to increased productivity, sustainability, and profitability in agriculture.

## 2. Methodology

The procedures used in the materials and techniques are summarised in Figure 1, and each step is covered in more detail in the sections that follow.

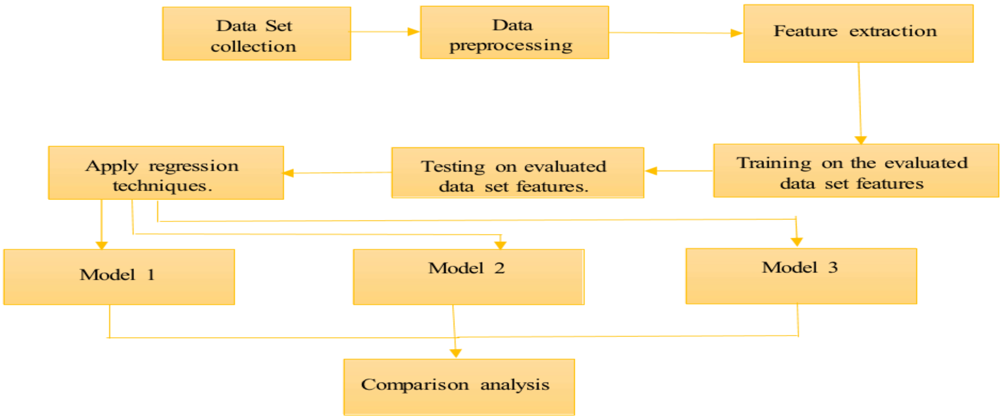
Data Set collection
Data preprocessing
Feature extraction
Training on the evaluated data set features
Testing on evaluated data set features.
Apply regression techniques.
Model 1
Model 2
Model 3
Comparison analysis

**Figure 1:** Block diagram of the proposed methodology

### 2.1 Data set availability

Predicting crop yields is a significant agricultural issue. For making judgements regarding agricultural risk management and generating forecasts for the future, it is vital to understand that agricultural productivity is primarily influenced by weather conditions (rain, temperature, etc.), pesticides, and reliable data regarding past crop yield. All the publicly accessible data used here is from the World Data Bank and the FAO (Food and Agriculture Organisation) [12-13]

### 2.2 Feature extraction and selection

Various features like area, item, year, domain, temperature, pesticides quantity, ha/hg_yield, rainfall are considered of various crops like potatoes, maize, wheat, rice, paddy, soyabeans, sorghum, sweet potatoes, cassava, yams. We discard area and item from the feature list as they were not providing any significant information.

### 2.2.3 Checking the data for outliers.

Examining a dataset to find any data points that differ noticeably from the rest of the data is known as "checking the data for outliers." Outliers are observations that stand out from the rest of the dataset unexpectedly and may point to mistakes, abnormalities, or significant insights [14]. Figure 2 represents the information of considered features w.r.t outlier's values. Observations that appear as extreme values, distant from the bulk of the data, or outside the expected range may be considered outliers.

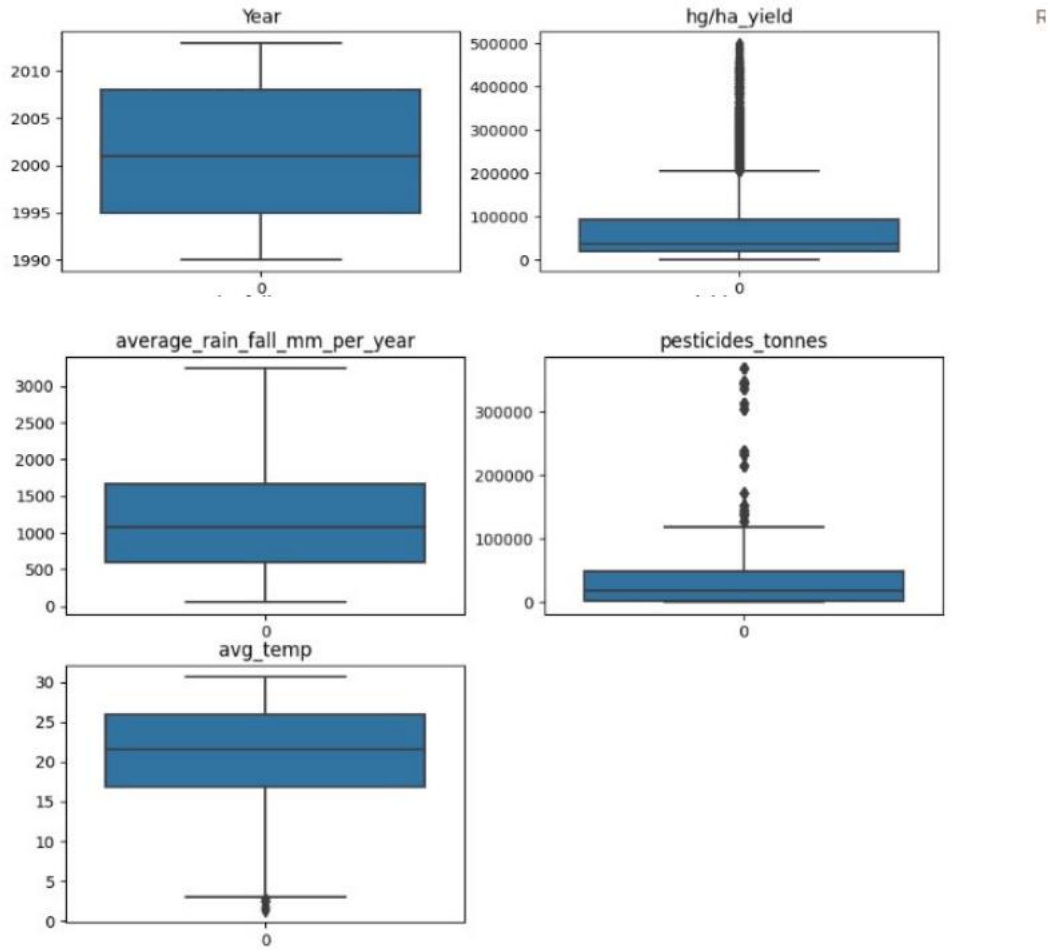

**Figure 2:** Outliers information

Heatmap in figure 3 also defines the correlation among extracted features. Heatmap uses a grid of coloured squares to symbolise values for a main variable of interest across two axis variables [15].

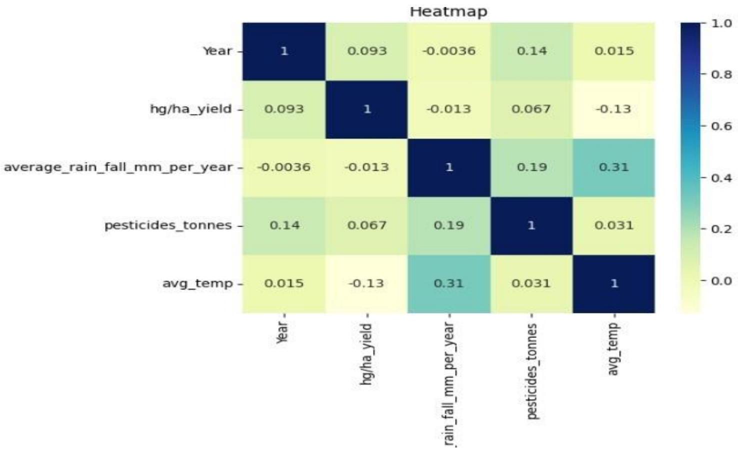

**Figure 3:** Heatmap generation for quantify data correlation.

### 2.3. Algorithms for machine learning are utilised in the testing and training phases.

#### a) Linear regression

Linear regression is a statistical modelling technique used to establish a linear relationship between a dependent variable and one or more independent variables [16]. In the context of crop yield prediction, linear regression can be applied to predict the yield of a crop based on various input factors such as weather conditions, soil characteristics, or management practices. Goal of linear regression is to fit a line or hyperplane that best represents the relationship between the independent variables and the dependent variable. The line or hyperplane is determined by estimating the coefficients (slopes) and intercept that minimize the differences between the predicted values and the actual values of the dependent variables [17]. In simple linear regression, there is only one independent variable, and the relationship can be represented by a straight line-equation of the form:

$Y = b0 + b1 * X$ (1)

Where, Y is the dependent variable, X is the independent variable, b0 is the intercept and the b1 is the slope. Multiple linear regression extends this concept to include multiple independent variables:

$Y = b0 + b1X1 + b2X2 + \ldots\ldots bnXn$ (2)

Where, Y is the dependent variables X1, X2,……Xn are the independent variables, b0 is the intercept, and b1,b2 ,……..bn are the slopes associated with each independent variables. To find the best-fitting line, linear regression uses a method called ordinary least squares estimation. This method calculates the vales of the coefficients that minimize the sum of the squared differences between the predicted values and the actual values. Once the coefficients are estimated, they can be used to make predictions for new data points based on the values of the independent variables.

#### b) Polynomial regression

Polynomial regression is a form of regression analysis that models the relationship between the dependent variable and the independent variable(s) as an nth degree polynomial [18]. It is an extension of linear regression, where the relationship is modelled as a straight line. In polynomial regression, the relationship between the dependent variable (Y) and the independent variable (X) is represented by an equation of the form:

$Y = b0 + b1X + b2X^{\wedge}2 + \ldots. bn * X^{\wedge}n$ (3)

Where b0, b1+….. bn are the coefficients, and n is the degree of the polynomial. Polynomial regression allows for more flexible and curved relationships between the variables. This enables the model to capture non-linear patterns that cannot be represented by a straight line. Similar to linear regression, polynomial regression aims to find the best fitting curve that minimizes the difference between the predicted values and the actual values of the dependent variable. The coefficients (b0, b1, ..., bn) are estimated using a method like ordinary least squares (OLS), which minimizes the sum of squared differences. The choice of the degree (n) in polynomial regression depends on the complexity of the relationship between the variables and the available data [19]. A higher degree polynomial can fit the training data more closely, but it may overfit and perform poorly on new, unseen data. Therefore, it's important to consider the balance between model complexity and generalization.

#### c) XGBoost

XGBoost named as Extreme Gradient Boosting, is an optimized implementation of gradient boosting, a machine learning algorithm that is widely used for both regression and classification tasks. XGBoost is known for its efficiency, speed, and performance in handling large-scale datasets [20]. Gradient boosting is an ensemble learning technique that combines multiple weak predictive models to create a stronger predictive model. It works by iteratively training new models that attempt to correct the mistakes of the previous models. The models are added to the ensemble in a way that minimizes the overall prediction error. XGBoost improves upon traditional gradient boosting by introducing several enhancements and optimization.

XGBoost includes regularization techniques to control model complexity and prevent overfitting. It incorporates both L1(Lasso) and L2 (Ridge) regularization terms to penalize large coefficients values. Its versatility, speed, and ability to handle complex datasets have made it a popular choice among data scientists and machine leaning practitioners. It has builtin handling for missing values in the data. It can automatically learn how to handle missing values in the data. It eliminates the need for explicit imputation.

## 2.4.Evaluation parameters

The effectiveness of a classifier can be assessed using a variety of evaluation criteria. Some of the most common parameters includes:

### a) Mean squared error:

How closely a regression line resembles a set of points using the mean squared error (MSE) accomplished by squaring the distances between the points and the regression line. The squaring is required to eliminate any unfavourable indications. Additionally, it emphasises bigger discrepancies [21]. The forecast is more accurate the lower the MSE.

$$MSE = \quad {}^{12} \quad \left(-\right) \qquad (4)$$

$$*actual - forecast\ n$$

Where, n defines as number of items or samples.

### b) R2 score:

It is called the coefficient of determination. It ranges from 0 to 1, quantifies how accurately a statistical model forecasts a result [22].

$$R2 = 1 - \underline{\qquad\qquad\qquad} \ Sum$$

$$of squares of residuals \qquad (5)$$

$$Totalsum\ of sqaures$$

## 2.5 Result analysis

The results of this study showed the comparison analysis of three regression models. Table 1 defines result analysis in terms of MSE and R2 score values. The learning outcomes of models differ in terms of their ability to capture non-linear relationships, handle interactions, and provide accurate predictions. Polynomial regression does not perform well as defined in table in terms of MSE and R2 score.

**Table 1.** Model performance

| Model name | MSE | R2 score |
|---|---|---|
| Linear regression | 9.699018E+08 | 7.327007E-01 |
| Polynomial regression | 1.825377e+31 | -5.030634e+21 |
| XGB | 5.192567e+08 | 8.568959e-01 |

Figure 4. represents the accuracy and the error rate of the linear, polynomial and XGB regression models.

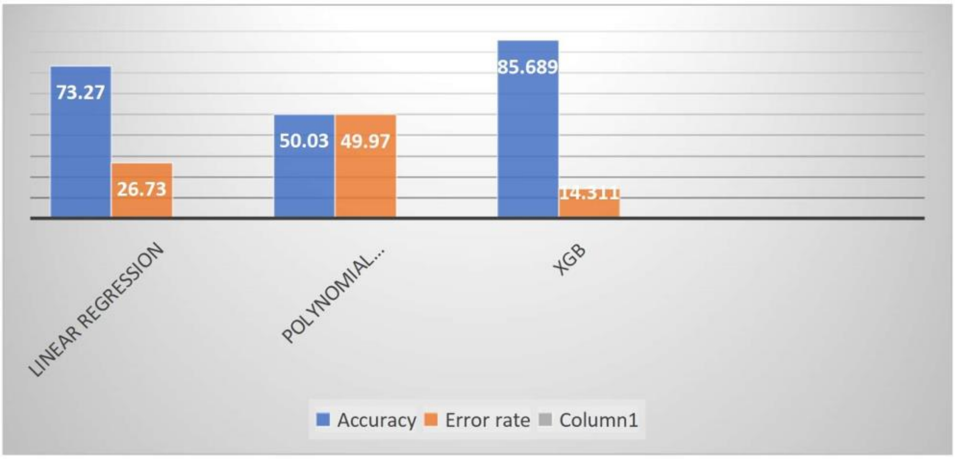

**Figure 4:** Analysis of accuracy and error rate

## 2.6 Discussion and conclusion

This research study's objective is to examine how machine learning techniques can be used to provide a precise crop yield prediction. In the presented research work, we used three regression models for predicting crop yield. As compares to other regression models, XGB performs better.

Images of the field and crop can be analysed using a variety of Artificial Intelligence (AI), Deep Learning (DL), and Computer Vision (CV) techniques to determine if they are infected with any diseases or the presence of weeds, which affect the quality of the crop, and can be separated from the healthy crops as soon as is practical.

## References

1. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.
2. Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.
3. Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021). An interaction regression model for crop yield prediction. Scientific reports, 11(1), 17754.
4. Zhang, L., Lei, L., & Yan, D. (2010, July). Comparison of two regression models for predicting crop yield. In 2010 IEEE International Geoscience and Remote Sensing Symposium (pp. 1521-1524).
5. Qaddourn, K. and EL. Hines, 2012. Reliable yield prediction with regress10n neural networks. Proceedings of the 12th WSEAS Intermtional Conference on Signal Processing, Computational Geometry and Artificial Intelligence, August 21-23, 2012, WSEAS Press, Turkey, Istanbul,-pp: I.
6. Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Attribute selection impact on linear and nonlinear regression models for crop yield prediction. The Scientific World Journal, 2014.
7. Kumar, R., Singh, M. P., Kumar, P., & Singh, J. P. (2015, May). Crop Selection Method to maximize crop yield rate using machine learning technique. In 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM) (pp. 138-145). IEEE.
8. Ngaruye, I.; von Rosen, D.; Singull, M. Crop yield estimation at district level for agricultural seasons 2014 in Rwanda. Afr. J. Appl. Stat. 2016, 3, 69–90.
9. Chakraborty, D.; Saha, S.; Sethy, B.K.; Singh, H.D.; Singh, N.; Sharma, R.; Chanu, A.N.; Walling, I.; Anal, P.R.; Chowdhury, S.; et al. Usability of the Weather Forecast for Tackling Climatic Variability and Its Effect on Maize Crop Yield in Northeastern Hill Region of India. Agronomy 2022, 12, 2529.
10. Beulah, R. (2019). A survey on different data mining techniques for crop yield prediction. Int. J. Comput. Sci. Eng, 7(1), 738-744.
11. Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. Computers and electronics in agriculture, 121, 57-65.
12. http://www.fao.org/home/en/
13. https://data.worldbank.org/
14. Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. International Journal of Computer Science Issues (IJCSI), 9(1), 307.
15. Tukey, J. W. (1977). Exploratory data analysis (Vol. 2, pp. 131-160).
16. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In An Introduction to Statistical Learning: with Applications in Python (pp. 69-134). Cham: Springer International Publishing.
17. Sellam, V., & Poovammal, E. (2016). Prediction of crop yield using regression analysis. Indian Journal of Science and Technology, 9(38), 1-5.
18. Shastry, A., Sanjay, H. A., & Bhanusree, E. (2017). Prediction of crop yield using regression techniques. International Journal of Soft Computing, 12(2), 96-102.
19. Heiberger, R. M., Neuwirth, E., Heiberger, R. M., & Neuwirth, E. (2009). Polynomial regression. R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics, 269-284.
20. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
21. Marmolin, H. (1986). Subjective MSE measures. IEEE transactions on systems, man, and cybernetics, 16(3), 486-489.
22. Lupón, J., Sanders-van Wijk, S., Januzzi, J. L., De Antonio, M., Gaggin, H. K., Pfisterer, M., ... & Bayes-Genis, A. (2016). Prediction of survival and magnitude of reverse remodeling using the ST2-R2 score in heart failure: a multicenter study. International journal of cardiology, 204, 242-247.