ISSN: 0974-5823 Vol. 7 No. 6 June, 2022

International Journal of Mechanical Engineering

Forecasting Crop yield using modified weighted ensmebling technique

Shivani S. Kale
Research Scholar, VTU. Assistant Professor ,Dept of CSE, ,KITCOEK, Kolhapur, Maharashtra
Email:shivanikale33@gmail.com
Dr. Preeti S. Patil.
Professor, Department of Information and Technology, D.Y.Patil College of Engineering ,Akurdi, Pune.MH
dr.preetipatil.dypa@gmail.com
Dr. Mamatha G.
Professor, Department of Information science and Engineering, RNS Institute of Technology, Bengaluru-560098.
Email: nidhimam.joshi2@gmail.com

I. Introduction

The paper details the research and application of various machine learning algorithms on a range of crops using factors such as temperature, rainfall, and various soil properties in about 392 districts across India. The purpose of this document is to advise farmers on what crop yields they can expect. As a result, the first stage is to create and implement a basic learner model, and then to assess performance. The ensemble technique is used in step 2 to measure the performance. Step 3: The optimized ensemble is used to calculate performance measures. All of the models are compared against each other based on various parameters. The paper is divided into four main sections. Section 1 covers datasets and data pre-processing. In Section II, the various ML techniques are discussed. Section III is devoted to a discussion of model performance using the acquired results. Finally, the findings in Section IV bring the paper to a conclusion.

II. Data set

State, District, year, season, crop, temperature, rainfall, production, N, P, K, Ph, Fc, Oc, and Zn are among the 15 features included in the dataset. Nearly 2 lakh records from 392 districts make up the collection. From 1997 to 2018, the data is available. The information was gathered from the government's website.

1. Data Preprocessing

Before feeding the dataset into the machine learning model, make sure it's correct. Datasets are subjected to the practice of eliminating erroneous, incomplete, and inaccurate data as well as substituting missing information.

2. Dataset Resources

The information is gathered from government websites. Weather data was available on the website "climateknowledgeportal.worldbank.org." "soilhealth.dac.gov.in" provided the data for soil parameters. "data.gov.in" was used to compile the agricultural production statistics.

3. Data Pre-processing

Steps used to pre-process data before it is fed into the model

Step 1 Get a dataset with suitable values for all rows containing null values.

Step 2 – Seasons are encoded using 0 and 1 value.

Step 3: The dataset is separated into crop-specific data frames, with unnamed column values being removed.

Step 4 – Using Lambda, the outlier is removed.

Step 5- Normalization of data

The normalisation equation is as follows:

$$Xnorm = X\text{-}Mean(column)/Std.Dev \quad \text{Equation (2)}$$

It returns a normalised version of X, with each feature's mean value set to 0 and the standard deviation set to 1.

Step 6- Splitting the dataset into two parts: a training set and a testing set. The data is split into two parts: 80:20.

Step 7: On the training set, K-fold cross validation is used with a value of K=5.

III. Performance Evaluation Metrices

I. Mean Absolute Error (MAE).

$$e_i=|yi\text{-}xi| \qquad \text{Equation (3)}$$

Where yi=Predicted Value, *x*i=Actual value and n=number of samples.

The MAE is calculated by taking the absolute errors' arithmetic average. If the MAE number is low, the model's performance is good.

II. Mean Square Error (MSE)

$$MSE = \sum (y_i\text{-}\tilde{y}_i)^2 \qquad \text{Equation (4)}$$

MSE is the square of the mean of the error, which is calculated by subtracting the projected value from the actual value for an n-sample population. It's the average of the errors' squares. The lower the MSE number, the better the model's performance.

III. Median Absolute Error (MAE)

Outliers are mostly dealt with via the Median Absolute Error. The loss is determined in MAE by computing the median of all absolute deviations with regard to the target and the prediction.

IV. Variance Score

The difference between actual data and a model is measured using explained variance. If the score is high, it indicates that the association is strong. As a result, a high variance score indicates that the model can generate more accurate predictions.

V. R Squared Score

It shows the link between the data and the regression line. If this score is 0, it suggests that none of the response data variability around the mean is present. If the R2 Score is 100%, it explains all of the variability in anticipated data around the mean.

IV. Base Learners Algorithms

Model 1: Linear Regression

Multiple input factors influence the quantitative response or prediction in multivariable linear regression. If there is a linear relationship between the independent and dependent variables, this model performs well.

$$Y = \beta' + \beta 1X1 + \beta 2X+ + \cdots + \beta\text{-}X\text{-} + \mathrm{e}$$

Equation (5)

The dependent variable is Y, the independent variables are X, the coefficients are $\beta 1$, and the error term is e..

Model 2: Lasso Regression

Least absolute shrinkage and selection operator is abbreviated as LASSO. By assigning zero values to their coefficients, the least contributing variables are removed in this strategy. It is a method of regularisation. A penalty term is employed in the LR LASSO model to decrease coefficients towards zero

Model 3: Decision Trees

Splitting the data set into smaller subsets to forecast the target value is the rationale of decision trees. The leaf node of a decision tree represents a condition, whereas the branches indicate the outcome of those conditions. Either the current rule for maximum depth is met, or no more gain can be obtained, the splitting comes to an end.

Model 4: Extreme Gradient Boosting (XGBoost)

The XGBoost method combines weak prediction models' predictions. It's a based on trees ensemble approach. XGBoost's predictions are based on learning from the mistakes of previous forecasters.

Model 5: Support Vector Machines with polynomial kernel

SVR is mostly used for classification, although it can also be utilised for regression. Non-linearity in the data is discovered and used to create a reliable prediction model.

Model 6: Random Forests

Random Forests is an ensemble approach in and of itself. During the training phase, the approach creates a forest of decision trees. Each tree's prediction is combined to anticipate the ultimate outcome.

V. Ensemble Techniques

Data science specialists employ ensemble learning, a powerful machine learning algorithm, across sectors. Ensemble learning is a technique that combines the predictions of numerous machine learning models into a single forecast**.**

The use of different models to combine judgments can help to improve overall performance. As a result, one of the most important reasons for employing ensemble models is to solve three problems: noise, bias, and variance. In such a case, if the ensemble model does not provide the collective experience to enhance the accuracy of prediction, then the base learners need to change.

1. Crop yield Prediction using Average Ensemble Technique

Ensemble technique is the approach where different base learners contribute to predict. In average ensemble Regression model, the final prediction of ensemble model is calculated by taking average of the base member predictions. Generally, for getting good result for ensemble model, variety of base learner is opted.

$$\text{Average Ensemble} = 1/j \sum_{j=0}^{5} wjxj, \quad \text{Equation (6)}$$

Where j is number of base models

In average ensemble model final prediction is equally dependent on all base models even if all are not efficiently contributing to the prediction. So, to get more accurate and improved results the weights can be optimized so that the bias and variance will be least.

2. Proposed modified optimized weighted ensemble Algorithm

Step 1: Initialization of weights

Step 2: Train Base models from M using training data X and Target output Y

Step 3: Defining Objective Function

$$J(\theta)=1/n\sum_{i=1}^{n} (Yi-\sum_{j=1}^{k} \theta_j \hat{y}_{ij})^2 + \lambda \sum_{j=1}^{k} (\theta_j)^2$$

Equation (7)

Step 4: To get the optimum weights for each model to contribute to final predictions the optimization problem will be solved as follows to find minimum of θ

$$\text{Min } J(\theta)=\min 1/n \sum_{i=1}^{n} (y_i-\sum_{j=1}^{k} \theta_j \hat{y}_{ij})^2 + \lambda \sum_{j=1}^{k} (\theta_j)^2$$

Equation (8)

$$\theta = \begin{bmatrix} \theta_{1,} \\ \theta_{2,} \\ \theta_{3,} \\ \theta_k \end{bmatrix}$$

Where $\theta_1, \ldots ,\theta_k$ are the weights for 6 base models

n=number of samples, yi= Actual output of ith sample,

ŷij = Predicted output of jth model for ith sample,

k=number of models, λ=It is regularization parameter used to control the selected values of θ by optimization algorithm.

Step 5: Applying Gradient Descent using Jacobian for calculating optimum weight

This is modified average ensemble model where the weights according to importance of each model are defined using optimization function to get improved result

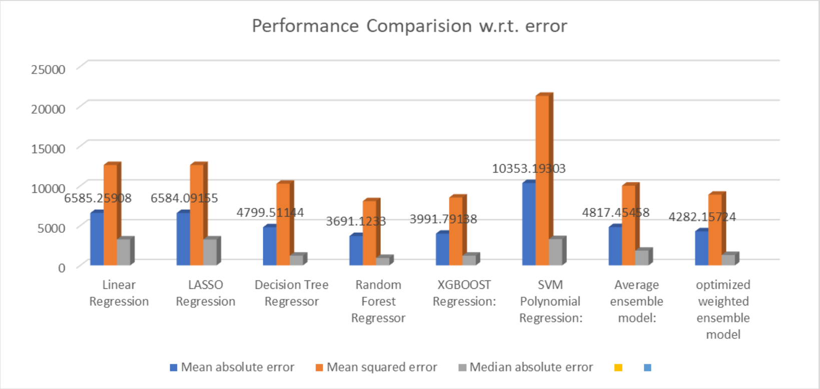
VI. Numerical Results

Weights column in Table 1 shows the weights calculated by proposed optimized weighted ensemble. Depending on the values of weight each base model contributes to final results. As the table shows the prediction capacity of SVM polynomial regression is not worth so its contribution will not be considered so weight value is 0 for the model.

Table 1 summarizes the performance of base ML models, Average ensemble model and proposed modified ensemble from the base ML models, Linear regression model makes the highest prediction error based on

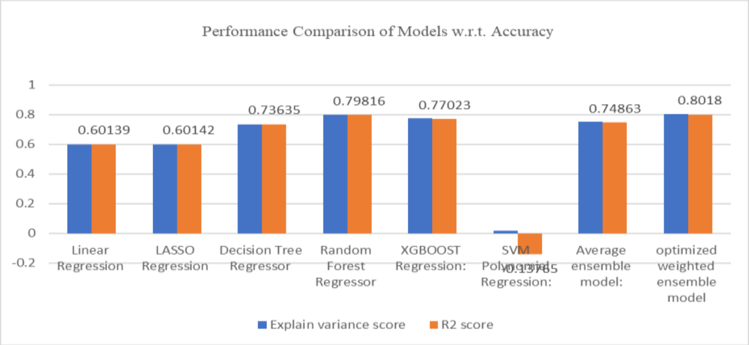
Mean absolute error, mean squared error and Median Absolute error. For accuracy score SVM Polynomial Regression shows lowest accuracy.

1. Performance Comparison of Models using Error Metric



Graph 1. Analysis of Models using Error Metrics

2. Performance Comparison of Models Using Accuracy Metrics

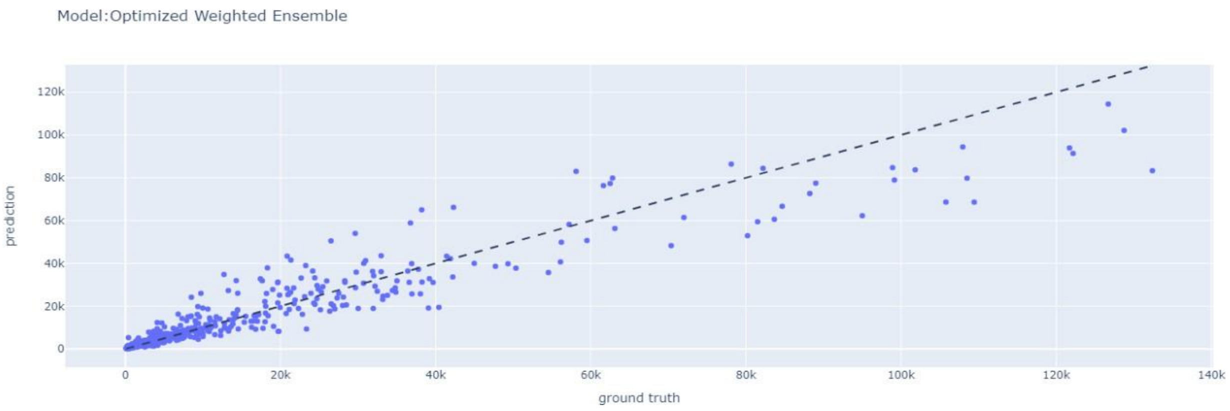


Graph 2. Analysis of Models using Accuracy Metrics

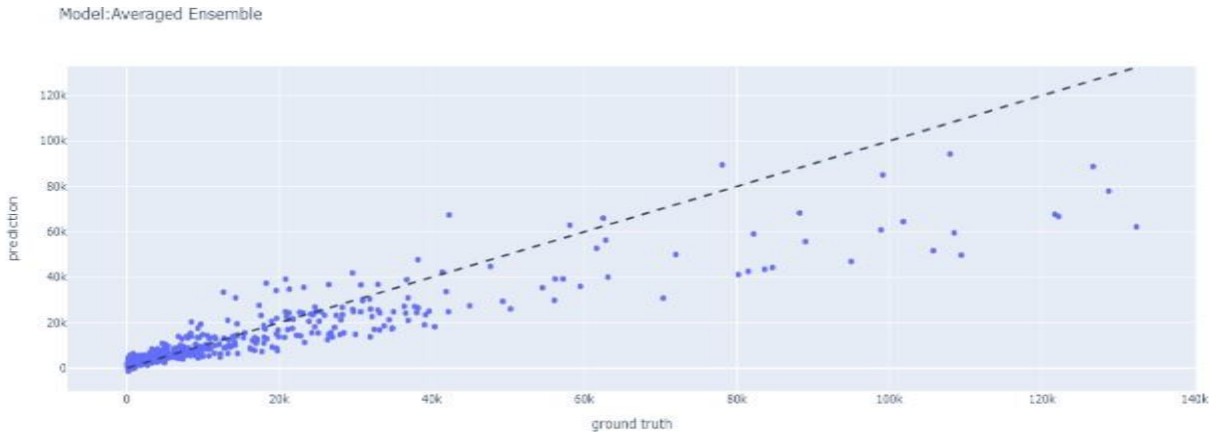
Table 1: Performance Evaluation for Crop: Groundnut

Models	Mean absolute error	Mean squared error	Median absolute error	Explain variance score	R2 score	**Weights**
Linear Regression	6585.25908	12614.15289	3265.03129	0.60141	0.60139	**0.1047**
LASSO Regression	6584.09155	12613.60861	3259.44364	0.60144	0.60142	**0.1035**
Decision Tree Regressor	4799.51144	10258.76247	1223.25697	0.73668	0.73635	**0.2326**
Random	3691.1233	8072.93787	937.58745	0.79854	0.79816	**0.2836**

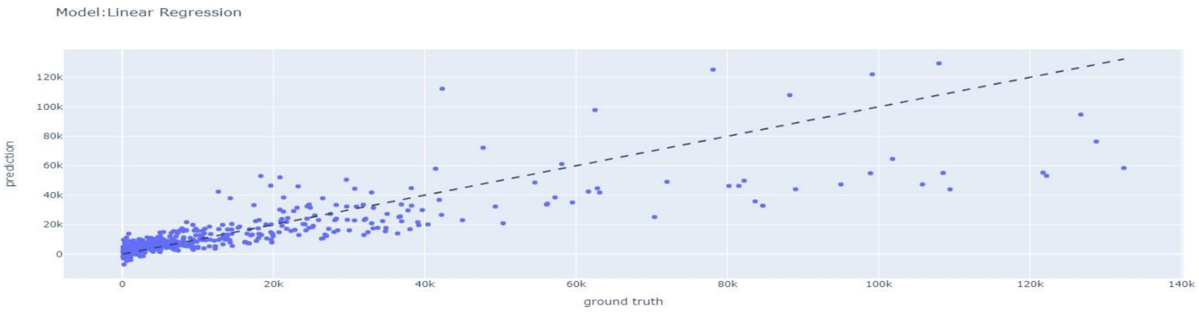
Forest Regressor						
XGBOOST Regression:	3991.79138	8527.623	1221.42504	0.77854	0.77023	**0.2757**
SVM Polynomial Regression:	10353.19303	21310.18539	3304.85026	0.01842	-0.13765	**0.0000**
Average ensemble model:	4817.45458	10017.13053	1848.94544	0.75168	0.74863	
optimized weighted ensemble model	**4282.15724**	**8892.98986**	**1312.0333**	**0.80212**	**0.8018**	



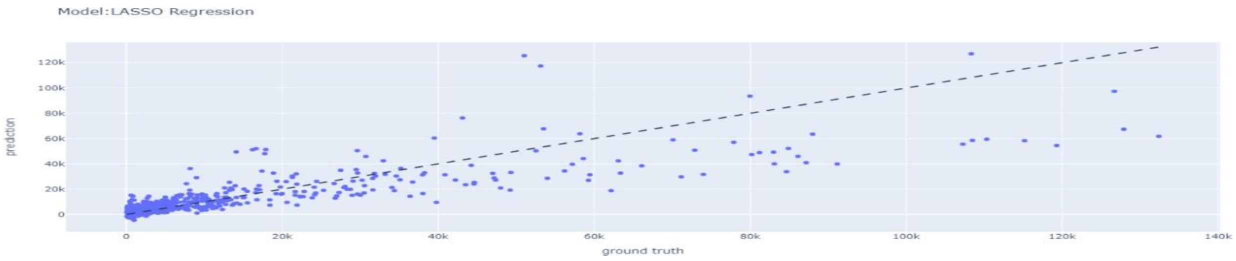
Graph 3. Proposed Optimized Weighted Ensemble



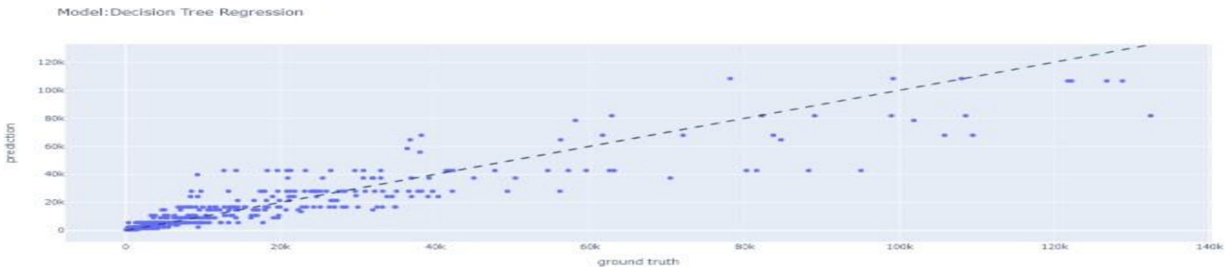
Graph 4. average Ensemble



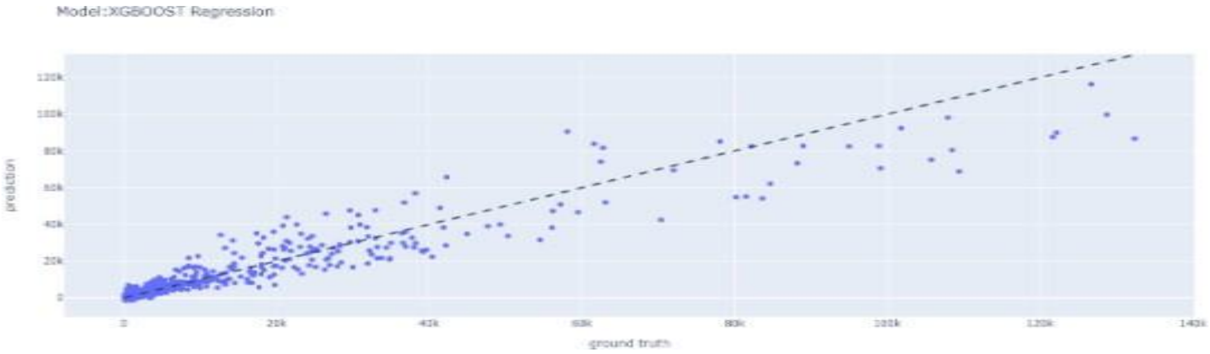
Graph 5. Linear regression Model Prediction



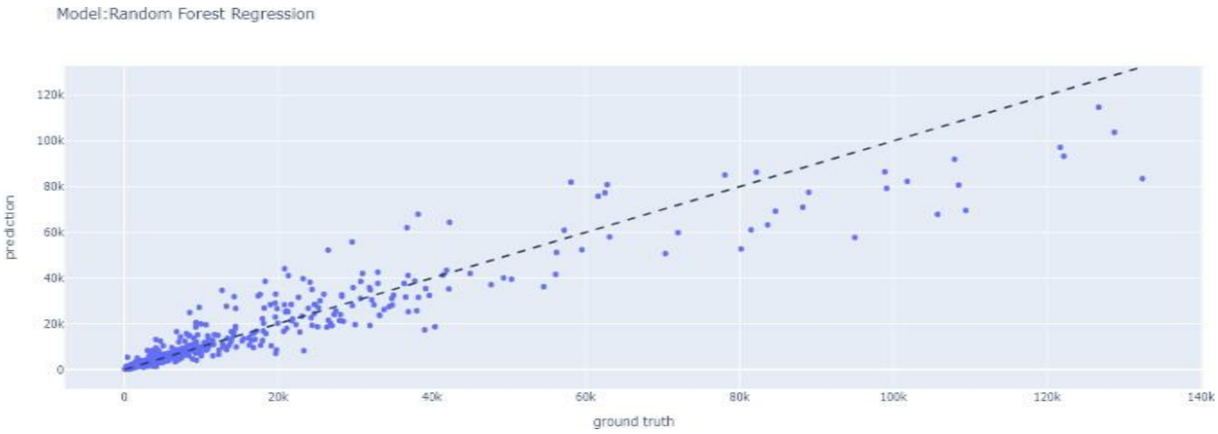
Graph 6. Lasso regression Model Prediction



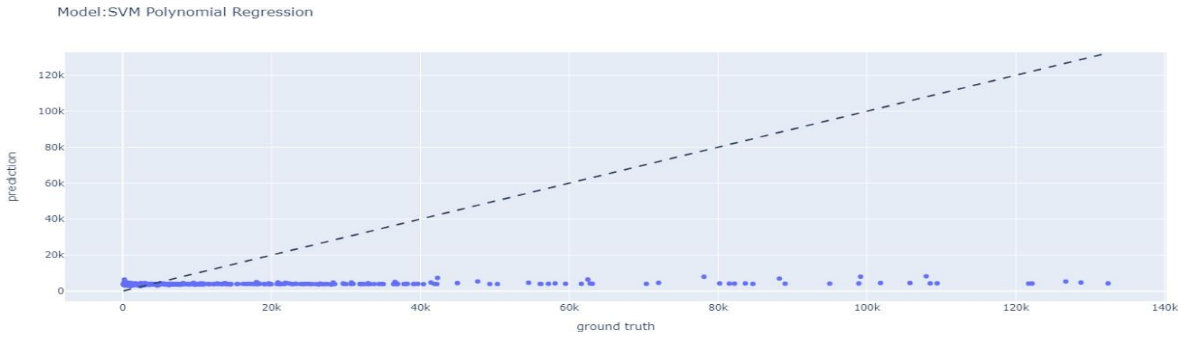
Graph 7. Decision tree regression Model Prediction



Graph 8. Decision tree regression Model Prediction



Graph 9. Decision tree regression Model Prediction



Graph 10. SVM Polynomial Regression

VII. Discussion

When compared to other models' graphs, the suggested modified weighted ensemble in graph 3 shows scattering of dots near the regression lines, indicating that the model attempted to predict with high accuracy and low error. As a result, the proposed model demonstrates when compared to the base models, there is a 20% gain in accuracy and a 30% reduction in variance.

VIII. Conclusion

Tentative prediction of yield will help the farmer regarding decision about cultivation of crops. Different base models are used to create ensemble model. The improved ensemble model gives better performance compared to unoptimized average ensemble. Comparison of results of different base models with proposed model is carried out in the study. The use of K-Fold cross validation help models to get more accurate results.

The proposed optimized weighted ensemble model that tries to balance both bias and variance of predictions and applies functions to discover the optimal weight to group multiple base models.

So, suggestion from the research work is that addition of more input parameters may improves the performance of model. Also selecting diverse base learners can contribute for better results to the ensemble model.

IX. REFERENCES

1. Mrs. Shivani S. Kale and Dr. Preeti Patil, “Data mining technology with fuzzy logic, neural networks and machine learning for agriculture”, at ASIC book series “Data management, analytics and innovation by springer 2018.
2. Mr. V. Sellam and E. Poovammal,” Prediction of crop yield using regression analysis”, Indian Journal of Science and technology,Vol 9 (38),October 2016.
3. Mrs. Shivani S. Kale and Dr. Preeti S. Patil, “Use of data mining technology in agriculture sustainable development”, IJCRT, Volume 6, Issue 2. April 2018.
4. Mrs. Shivani S. Kale and Dr. Preeti S. Patil “Application of modified ensemble technique using weights optimization for Crop Yield Prediction”, GIS SCIENCE JOURNAL, ISSN NO : 1869-9391, VOLUME 7, ISSUE 12, 2020
5. Mohsen Shahhosseini1, Guiping Hu1*, Sotirios V. Archontoulis, “Forecasting Corn Yield with Machine Learning Ensembles”